

# 1 A synthetic dataset for Time Series Super-Resolution 2 with Deep Learning

3 Julio Ibarra-Fiallo<sup>1</sup>, Juan A. Lara<sup>2</sup>, and D'hamar Agudelo-Moreno<sup>1</sup>

4 <sup>1</sup>Colegio de Ciencias e Ingenierías, Universidad San Francisco de Quito, Cumbayá, Ecuador

5 <sup>2</sup>Universidad de Córdoba, Córdoba, España

6 \*corresponding author: Julio Ibarra-Fiallo (jibarra@usfq.edu.ec)

## 7 ABSTRACT

The increasing application of temporal signal analysis in fields like biomedical engineering, telecommunications, and industrial monitoring emphasizes the need for high-quality data to train and evaluate advanced machine learning models.<sup>REV</sup> The increasing application of time-series analysis in fields like biomedical engineering, telecommunications, and industrial monitoring emphasizes the need for high-quality data to train and evaluate advanced machine learning models.<sup>REV</sup> Acquiring real-world temporal data at suitable resolutions is often limited by ethical, economic, or practical constraints. To address this, we introduce CoSiBD (Complex Signal Benchmark Dataset for Super-Resolution), a synthetic dataset of complex temporal signals designed for training and assessing AI models, particularly deep learning systems, in tasks like temporal super-resolution and signal processing. CoSiBD comprises 2,500 high-resolution signals ( $N = 5,000$  samples each over a reference domain  $\tau \in [0, 4\pi]$ ) with corresponding low-resolution versions at four levels (150, 250, 500, and 1,000 samples) obtained via simple uniform decimation (uniform subsampling) of the original sequence.<sup>REV</sup> Each signal is provided in three formats (NumPy arrays, plain text, and JSON) with comprehensive metadata documenting all generation parameters, including random seeds for full reproducibility. Each signal is provided in three formats (NumPy arrays, plain text, and JSON) with comprehensive metadata documenting all generation parameters for reproducibility.<sup>REV</sup> CoSiBD includes diverse signals with non-uniform frequency modulations, capturing gradual transitions and abrupt high-frequency events to mirror real-world dynamics to approximate a range of dynamics observed in practice.<sup>REV</sup> The dataset provides both clean and noisy variants. The dataset provides both clean and noisy high-resolution signals; additionally, multiple subsampled versions are provided to support SR benchmarking.<sup>REV</sup> Low-resolution signals are provided at four target resolutions (150, 250, 500, and 1,000 samples). Subsampling is performed using two approaches: direct re-evaluation at lower time resolutions and uniform decimation.<sup>REV</sup> The dataset is generated by combining distinct frequency bands, non-uniform intervals, and probabilistic frequency assignments to create realistic patterns, with smoothing achieved through spline interpolation. We report a technical validation focusing on spectral consistency across sampling rates and noise; CoSiBD supports training and evaluation. Validated for spectral consistency across sampling rates and noise, CoSiBD supports training and evaluation.<sup>REV</sup>

## 9 Background & Summary

10 The analysis and simulation of temporal signals are fundamental across science and engineering, supporting insights into  
11 dynamic processes.<sup>REV</sup> The analysis and simulation of temporal signals are fundamental across science and engineering. These  
12 techniques provide critical insights into dynamic processes in multiple domains.<sup>REV</sup> In biomedical research<sup>1</sup>, electroencephalography  
13 (EEG) and electrocardiography (ECG) analyses reveal brain and heart function<sup>2,3</sup>. Telecommunications rely on signal  
14 processing to ensure data fidelity across noisy media<sup>4</sup>, while finance uses time-series forecasting for risk and trend analysis<sup>5</sup>.  
15 Industrial monitoring detects equipment faults using temporal patterns<sup>6</sup>, and environmental science applies similar techniques  
16 to climate and environmental monitoring using remote-sensor time series<sup>7</sup>. Developing<sup>REV</sup> tools for interpreting time-varying data continues to support both scientific discovery and practical applications.

17 Recent advances in deep learning have contributed significantly to this field by enabling automatic extraction of complex  
18 features from raw signals. Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), including Long  
19 Short-Term Memory (LSTM) units, and Generative Adversarial Networks (GANs) have demonstrated improved performance  
20 over traditional techniques in image, speech, and time-series processing tasks<sup>8,9</sup>. These models support fine-grained signal  
21 reconstruction and forecasting, allowing researchers to explore temporal dynamics in new ways.

22 Despite this progress, deep learning methods for temporal signal processing often require large quantities of labeled, high-  
23 quality data. Access to such data is frequently constrained by medical privacy regulations such as the General Data Protection  
24 Regulation (GDPR) and the Health Insurance Portability and Accountability Act (HIPAA)<sup>10</sup>. In other domains, including

28 environmental monitoring using remote sensors and industrial monitoring<sup>remote sensing and industrial monitoring</sup><sup>REV</sup>, data  
29 availability is limited by practical and economic barriers to sensor deployment and data collection<sup>5</sup>. These limitations are partic-  
30 ularly relevant in super-resolution (SR) tasks, where models require paired low- and high-resolution signals for effective training.

31  
32 Temporal SR, which enhances resolution over time, has broad potential. In biomedical monitoring and sensing, SR can  
33 help reconstruct higher-resolution physiological time series (e.g., ECG/EEG), potentially improving the analysis of neural oscil-  
34 lations<sup>2</sup> and subtle physiological irregularities<sup>3</sup>. SR also applies to audio/speech enhancement, industrial vibration monitoring,  
35 and telecommunications, where higher temporal resolution can increase sensitivity to rapid changes and improve signal quality.  
36 In medicine, for instance, it improves magnetic resonance imaging (MRI) and computed tomography (CT) scans, supporting  
37 earlier disease detection<sup>11</sup>. For EEG analysis, SR may help recover high-frequency components that aid in the study of neural  
38 oscillations<sup>2</sup> or detect subtle physiological irregularities<sup>3</sup>. In remote sensing, SR helps refine satellite imagery<sup>7</sup>, while in  
39 telecommunications it contributes to enhanced signal reliability. It also has applications in industrial monitoring by increasing  
40 sensitivity to system changes.<sup>REV</sup>

41  
42 Traditional SR methods such as polynomial interpolation, frequency-domain transforms, and splines each have limitations.  
43 Polynomial models are often insufficient for capturing nonlinear dynamics; frequency-domain methods are susceptible to  
44 noise<sup>7</sup>; and splines, though flexible, may not generalize well to complex signal variability<sup>12,13</sup>. Many of these methods also  
45 assume uniform partitioning, which may not align with the multi-scale, irregular structure of natural temporal phenomena.

46 Deep learning offers adaptive alternatives to these traditional methods. CNNs are capable of modeling spatio-temporal  
47 structure, RNNs and LSTMs capture long-range dependencies in time, and GANs can learn high-resolution representations  
48 through adversarial training<sup>8,9</sup>. While GANs have achieved strong results in image SR<sup>14</sup>, their application to time-series SR  
49 remains relatively new. Preliminary work on synthetic time-series generation indicates potential<sup>14,15</sup>, but the lack of accessible,  
50 high-quality paired datasets remains a significant barrier to progress.

51  
52 Synthetic datasets offer one solution to this problem, allowing researchers to design reproducible training environments  
53 that reflect the structure and variability of real-world signals. Prior studies have used synthetic data in domains such as fluid  
54 dynamics<sup>16</sup>, bioimaging<sup>17</sup>, and live-cell imaging<sup>18</sup>, demonstrating that synthetic approaches can help simulate complexity  
55 while avoiding legal and practical restrictions associated with real-world data.

56  
57 To support research in super-resolution for time-series data, we present the Complex Signal Benchmark Dataset (CoSiBD).  
58 CoSiBD is a synthetic dataset composed of time-series signals with variable resolution, frequency characteristics, and noise  
59 levels. The dataset is intended to provide a resource for training and evaluating SR models under controlled, reproducible  
60 conditions. It includes non-stationary, piecewise-structured signals (via non-uniform interval partitioning with change-points),  
61 multiple levels of resolution and noise, a technical validation suite, and publicly available Python code to facilitate use. CoSiBD  
62 has been used in research presented at the International Conference on Signal Processing and Machine Learning<sup>15</sup> and is made  
63 available to support further development in deep learning approaches for temporal super-resolution.

64 To further position CoSiBD with respect to existing public synthetic time-series resources, we summarize representative  
65 datasets and simulators and highlight the practical gap addressed by our benchmark.<sup>REV</sup>

## 68 Related synthetic time-series resources

69 Publicly available synthetic resources for temporal signals exist, but they are typically designed for tasks other than time-series  
70 super-resolution (SR), or they target a specific domain. In wireless communications, the RadioML family provides large  
71 collections of synthetic complex I/Q sequences with varying SNR and channel impairments, mainly to benchmark automatic  
72 modulation classification rather than paired SR reconstruction<sup>19–21</sup>. In biomedical signal processing, physiological simulators  
73 such as ECGSYN (ECG) and SEREEGA (EEG) enable controlled generation with tunable morphology, sampling settings,  
74 and noise, supporting method development when real data access is constrained<sup>22–24</sup>. In power systems, LoadGAN provides  
75 multi-resolution generation of load time series across sampling rates and time horizons (from sub-second to long-term scales),  
76 but it is not distributed as a standardized paired SR benchmark<sup>25</sup>. Domain-specific paired low-/high-resolution training data can  
77 also be produced via physical forward modeling, e.g., low- and high-resolution 1D seismic traces for learning-based resolution  
78 enhancement<sup>26</sup>.<sup>REV</sup>

79 Table 1 summarizes these representative resources and highlights a practical gap: while many tools provide synthetic  
80 signals, they usually do not jointly offer (i) multi-factor paired LR–HR signals for time-series SR, (ii) a clear pairing protocol  
81 for low-resolution observations aligned to reconstructing the original HR target (here implemented via simple uniform decima-  
82 tion), and (iii) per-signal metadata enabling deterministic regeneration and principled benchmarking. CoSiBD is designed to

[REV 1]  
Added a short related work subsection and comparison table to position CoSiBD against representative publicly available synthetic time-series datasets and simulators and explicitly state the practical gap addressed (reviewer request: contextualize vs. existing resources).

Resource	Domain	Form	Paired LR–HR SR	Multi-resolution	Noise / artifacts	Reproducibility granularity
CoSiBD (this work)	Generic time series (complex-structured signals)	Dataset generator	Yes (LR → HR targets)	Yes (150/250/500/1000 → 5000)	Gaussian + structured interference; primary benchmark uses direct decimation	Per-signal metadata; deterministic regeneration (seed-controlled)
RadioML 2016.10A <sup>19,20</sup>	Wireless communications (I/Q)	Dataset	No (classification benchmark)	N/A (not SR)	Variable SNR + channel impairments	Dataset-level (labels/SNR); not per-sample “recipe”
RadioML 2018.01A <sup>21</sup>	Wireless communications (I/Q)	Dataset	No (classification benchmark)	N/A (not SR)	Simulated channel effects + SNR variability	Dataset-level; not SR-paired
ECGSYN <sup>22,23</sup>	ECG (physiology)	Simulator/tool	Configurable <sup>1</sup>	Configurable (via sampling settings)	Model-based; supports controlled variability	Configurable via simulator parameters (user-defined)
SEREEGA <sup>24</sup>	EEG (physiology)	Simulator/toolbox	Configurable <sup>1</sup>	Configurable (user-defined)	Supports noise and event-related components	Configurable via simulator parameters (user-defined)
LoadGAN <sup>25</sup>	Power systems load time series	Generator/tool	No (generation)	Yes (variable sampling rates)	Domain-specific variability (load patterns)	Tool-based; generation is configurable
Synthetic LR–HR seismic traces (example) <sup>26</sup>	Seismic traces (geophysics)	Paper-specific paired data	Yes (LR–HR pairs)	Typically limited (study-specific)	Study-dependent	Paired data available for the study; limited generality

**Table 1.** Representative publicly available synthetic time-series datasets and simulators related to signal processing and learning. “Form” indicates whether the resource is distributed primarily as a fixed dataset or as a simulator/generator. “Reproducibility granularity” summarizes whether exact per-sample regeneration is supported via documented parameters and seeds.

address this gap by providing multi-resolution paired signals, explicit nuisance modeling (noise and structured interference), and comprehensive metadata for reproducible SR benchmarking across multiple difficulty levels. Table 1 summarizes these representative resources and highlights a practical gap: while many tools provide synthetic signals, they usually do not jointly offer (i) multi-factor paired LR–HR signals for time-series SR, (ii) a clear pairing protocol for constructing low-resolution observations for SR, and (iii) per-signal metadata (including random seeds) enabling exact sample-level reproducibility. CoSiBD is designed to address this gap by providing multi-resolution paired signals, explicit nuisance modeling (noise and structured interference), and comprehensive metadata for reproducible SR benchmarking across multiple difficulty levels.<sup>REV</sup>

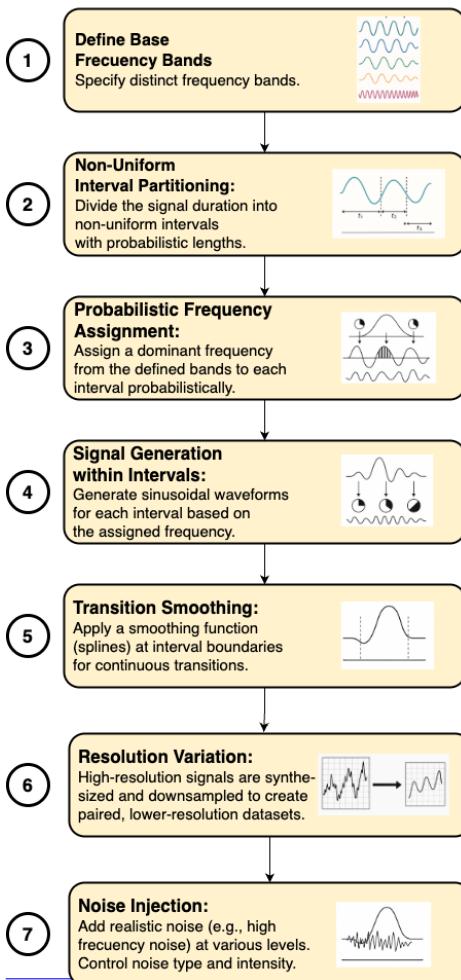
## Methods

The methodology used to generate the synthetic temporal signals that constitute the CoSiBD dataset is illustrated in Figure 1. The process was designed to produce signals that reflect general characteristics of real-world temporal data, such as variable frequency content, continuous transitions, and intermittent high-frequency activity. A key aspect of the procedure is the ability to produce signals at different resolution levels, supporting the generation of paired datasets for evaluating super-resolution (SR) algorithms.

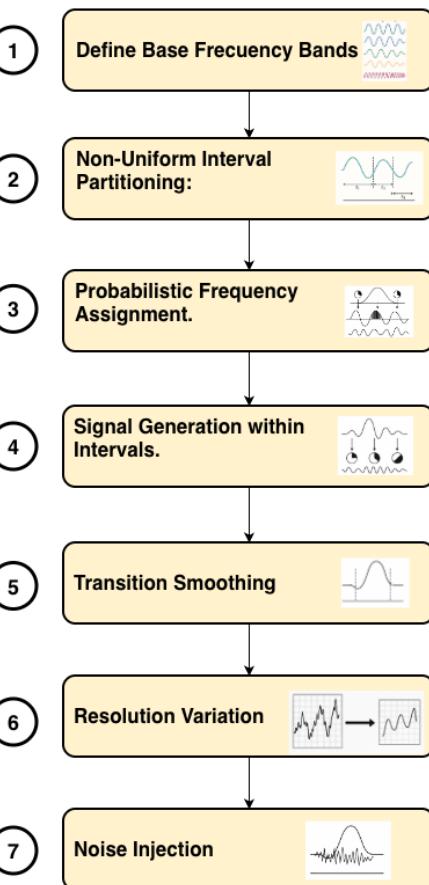
**Design rationale inspired by real signals.** To address the concern that the dataset is “too artificial”, we derived the simulator degrees of freedom from qualitative observations across representative physiological (EEG/ECG) and speech signals. In particular, real signals exhibit (i) non-stationary regime changes, (ii) coexisting low- and high-frequency components with intermittent transients, (iii) smooth amplitude-envelope evolution, and (iv) slow baseline drift and measurement noise. CoSiBD instantiates these properties via non-uniform interval partitioning with change-points, separate low/high-frequency bands,

<sup>1</sup>“Configurable” indicates that LR–HR pairs can be constructed by running the simulator at different sampling settings and/or applying controlled downsampling, but a standardized paired SR benchmark (multi-factor LR versions aligned to a fixed HR target) is not typically distributed as part of the resource.

## CoSiBD Dataset Generation Process

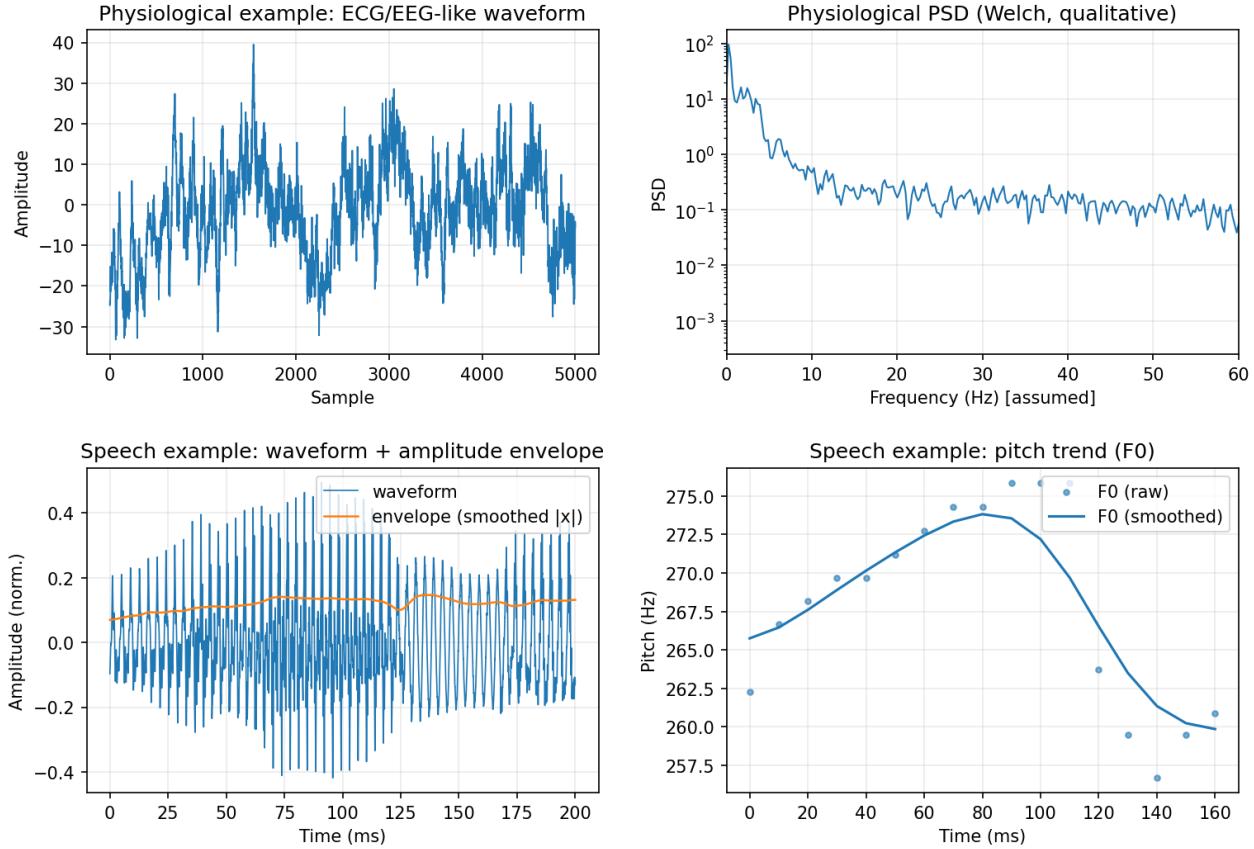


## CoSiBD Dataset Generation Process



**Figure 1.** Schematic overview of the CoSiBD signal generation process.

Real-signal properties motivating CoSiBD design (qualitative examples)

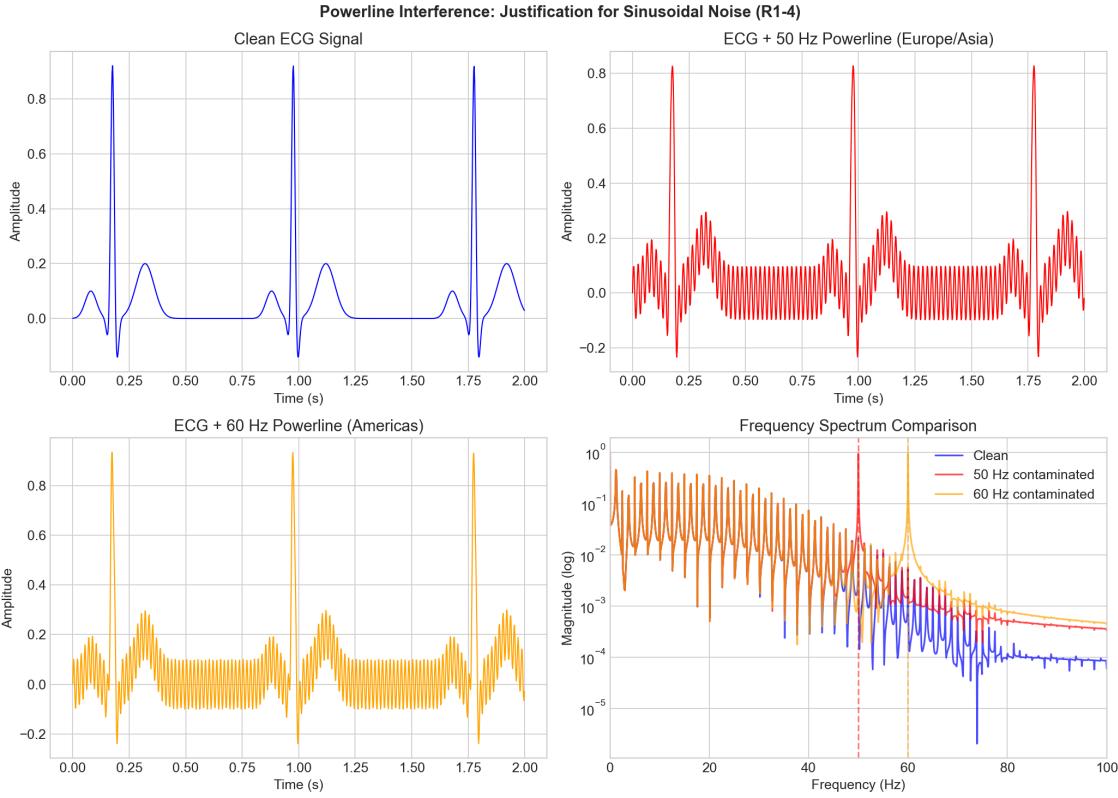


**Figure 2.** Qualitative real-signal properties motivating the CoSiBD design. The physiological example illustrates non-stationarity in the waveform and structured spectral content; the speech example illustrates amplitude-envelope dynamics and a smoothly varying pitch (F0) trend. These observations motivate CoSiBD mechanisms such as regime partitioning with change-points, low/high-frequency bands, and spline-based envelopes/frequency profiles.

101 spline-based envelopes and frequency profiles, and explicit offset/noise terms. Figure 2 provides qualitative examples of these  
 102 motivating properties; the goal is to capture challenging structure for SR benchmarking rather than match a specific domain  
 103 distribution.<sup>REV</sup>

104 The signal generation pipeline involves the following steps:

- 105 **1. Base frequency band definition:** A set of distinct frequency bands is defined to represent the underlying spectral content  
 106 of the signals. These can be adjusted to reflect application-specific characteristics.
- 107 **2. Non-uniform interval partitioning:** The total signal duration is divided into multiple intervals of variable length. The  
 108 interval lengths are determined probabilistically to introduce variability in the signal structure.
- 109 **3. Frequency assignment:** Each interval is assigned a dominant frequency band, sampled according to a predefined  
 110 probability distribution. This introduces spectral variation over time.
- 111 **4. Signal synthesis:** A sinusoidal waveform, or a combination of sinusoids within the assigned frequency band, is generated  
 112 for each interval. Signal parameters such as amplitude and phase are configurable.
- 113 **5. Transition smoothing:** To avoid discontinuities at interval boundaries, a smoothing function is applied to overlapping  
 114 segments. This ensures gradual transitions between intervals with different frequency content.
- 115 **6. Resolution variation:** All signals are initially synthesized at a high temporal resolution (5,000 samples over the domain  
 116  $[0, 4\pi]$ ). Lower-resolution versions are created using simple decimation (uniform subsampling). This keeps the



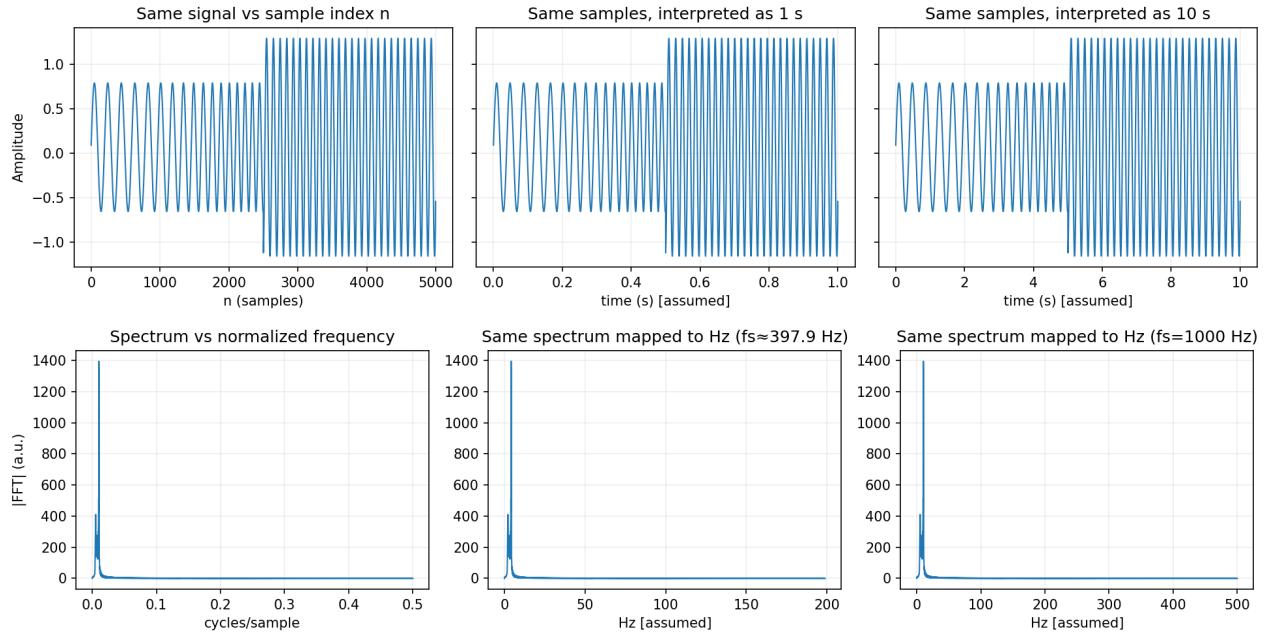
**Figure 3.** Qualitative motivation for the structured interference term used in CoSiBD. An illustrative example shows how adding a narrow-band sinusoidal component (interpretable as 50/60 Hz under the illustrative convention  $T = 4\pi$  s) produces the characteristic periodic contamination observed in real recordings, while broadband noise captures the measurement floor.

SR task aligned with reconstructing the original high-resolution target; the low-resolution observation is obtained by subsampling the original sequence without pre-filtering. Reconstructing low-pass filtered signals is not an objective of CoSiBD.<sup>REV</sup> using two distinct approaches: (1) direct re-evaluation by computing the signal at fewer time points using the original generation parameters, and (2) direct decimation (uniform subsampling).<sup>REV</sup> For reproducibility, given a high-resolution sequence  $x_{HR}[n]$  of length  $N = 5000$  and a target low-resolution length  $M \in \{1000, 500, 250, 150\}$ , we form  $x_{LR}[i] = x_{HR}[n_i]$  using the fixed index set  $n_i = \left\lfloor \frac{i(N-1)}{M-1} + 0.5 \right\rfloor$  for  $i = 0, \dots, M-1$  (applied identically to the time array). This reduces to standard stride decimation when  $M$  divides  $N$ .<sup>REV</sup>

7. **Noise injection:** Controlled levels of synthetic noise are added to the signals to emulate different data acquisition scenarios. Two noise types are implemented: Gaussian noise with configurable standard deviation (relative to signal amplitude) and structured sinusoidal noise bursts (deterministic sinusoidal components). Noise is applied probabilistically with 50% probability per signal. Two noise types are implemented: Gaussian noise with configurable standard deviation (relative to signal RMS) and structured sinusoidal interference (deterministic narrow-band components). Noise is applied probabilistically with 80% probability per signal; when noise is present, Gaussian noise is selected with 70% probability.<sup>REV</sup> Both the type and intensity of the noise can be configured.

**Rationale for structured 50/60 Hz interference and noise.** Real measurement pipelines frequently contain narrow-band interference (e.g., mains hum) superimposed on broadband sensor noise. To reflect this common acquisition artifact, CoSiBD includes an optional structured sinusoidal component in addition to Gaussian noise. CoSiBD signals are generated over a reference domain (by default  $\tau \in [0, 4\pi]$ ); interpreting  $\tau$  as physical time (and therefore reporting frequencies in Hz) requires an explicit time scaling. Throughout this manuscript we adopt an illustrative convention that maps the reference domain to a duration  $T = 4\pi$  seconds, under which the structured component can be interpreted as a 50/60 Hz-like powerline interference term, while the broadband term represents the measurement noise floor. Figure 3 illustrates this qualitative motivation; the intent is not to reproduce a specific device transfer function but to include realistic nuisance factors that SR models must handle.<sup>REV</sup>

**Sampling units and frequency interpretation.** CoSiBD signals are provided as discrete sequences  $x[n]$  (e.g.,  $N = 5,000$



**Figure 4.** Sampling/unit convention in CoSiBD. Top: the same discrete sequence  $x[n]$  can be plotted against the sample index or under different assumed time scalings. Bottom: the intrinsic frequency axis is normalized (cycles/sample); mapping to “Hz” depends on the assumed sampling rate  $f_s$  (two example mappings shown).

samples) that are directly used as inputs/targets by SR models. The internal generation domain  $\tau \in [0, 4\pi]$  is a reference parameterization; interpreting it as physical time requires choosing a duration  $T$  (in seconds) for the reference interval. Under this convention, the implied sampling rate is  $f_s = N/T$  and all frequencies reported in Hz scale linearly with  $4\pi/T$ . Throughout this manuscript, when reporting example frequencies in Hz we adopt the illustrative convention  $T = 4\pi$  s, yielding  $f_s \approx 5000/(4\pi) \approx 398$  Hz; other equally valid mappings exist depending on application. Consequently, any band-specific interpretation in Hz (e.g., “low/high” frequency ranges) should be understood under the chosen  $T$ ; changing  $T$  rescales all reported Hz values while preserving the underlying discrete sequences, which is a key feature of CoSiBD’s reference-domain design. Figure 4 illustrates that the discrete samples are unchanged under different time scalings and that Hz axes shift with the assumed  $f_s$ , while the normalized spectrum (cycles/sample) is invariant.<sup>REV</sup>

The parameters that govern each step of the generation process—such as interval length distributions, frequency band selection probabilities, smoothing function characteristics, sampling rates, and noise settings—can be configured to produce signal sets tailored to different domains or experimental conditions. All generation parameters, including random seeds, are documented in comprehensive metadata (`signals_metadata.json`), enabling exact reproduction of individual signals or the complete dataset. The generation pipeline is implemented in modular Python code available in the `SignalBuilderC` package, with clear interfaces for customization and extension. All generation parameters are documented in comprehensive metadata (`signals_metadata.json`), enabling deterministic reproduction of the official dataset release. In particular, the full dataset can be regenerated by rerunning the generator with a fixed global seed (`seed=42`). The generation pipeline is implemented in modular Python code available in the `SignalBuilderC` package, with clear interfaces for customization and extension.<sup>REV</sup> These configurations are included in the dataset’s accompanying code to support reproducibility and allow users to regenerate the signals under consistent conditions.

## 160 Data Records

161 The Complex Signal Benchmark Dataset (CoSiBD) is publicly available on Zenodo<sup>27</sup> and consists of synthetic temporal signals  
 162 created to support the development and evaluation of temporal super-resolution (SR) algorithms. This section provides an  
 163 overview of the dataset structure, content, and storage format.

164  
 165 The dataset includes a total of 7,800 signal samples divided into two main categories:<sup>REV</sup> The dataset comprises 2,500  
 166 high-resolution signals, each with corresponding subsampled versions at four resolution levels, organized into three main

167 categories:<sup>REV</sup>

- **High-resolution signals:** 2,500 signals with 5,000 samples each, spanning the domain  $[0, 4\pi]$ . Each signal is stored in three formats: NumPy compressed format (.npz), plain text (.txt), and JSON (.json). Per-signal metadata (frequency profiles with explicit change-points (`base_points` and `high_freq_points`) and segment labels (`variation_type`), amplitude envelopes, spline parameters, vertical offsets, noise configurations, and random seeds) is provided in a consolidated JSON file (`signals_metadata.json`) with one entry per signal, enabling exact regeneration.<sup>REV</sup> **High-resolution signals:** 2,500 signals with 5,000 samples each, spanning the domain  $[0, 4\pi]$ . Each signal is stored in three formats: NumPy compressed format (.npz), plain text (.txt), and JSON (.json). Per-signal metadata (frequency profiles with explicit change-points (`base_points` and `high_freq_points`) and segment labels (`variation_type`), amplitude envelopes, spline parameters, vertical offsets, and noise configurations) is provided in a consolidated JSON file (`signals_metadata.json`) with one entry per signal. The official dataset release can be deterministically regenerated using the fixed seed (`seed=42`).<sup>REV</sup>
- **Low-resolution signals,** obtained through controlled downsampling of the high-resolution versions, available at three distinct resolution levels.<sup>REV</sup> **Simple subsampled signals:** Uniform decimation (uniform subsampling) of each signal to four target resolutions (150, 250, 500, and 1,000 samples). These low-resolution versions serve as inputs for SR benchmarking against the original 5,000-sample target. Stored in .npz, .txt, and .json formats.<sup>REV</sup> **Simple subsampled signals:** Re-evaluation of each signal at four target resolutions (150, 250, 500, and 1,000 samples) using the original generation parameters. Noise is not re-applied in these re-evaluated versions (clean re-evaluation).<sup>REV</sup> Stored in .npz, .txt, and .json formats.<sup>REV</sup>

186 Noise is applied to both high-and low-resolution signals at different signal-to-noise ratio (SNR) levels (20 dB, 10 dB, and 5  
187 dB), integrated directly into the signal files.<sup>REV</sup> Reproducibility is ensured through documented random seeds: each high-  
188 resolution signal is generated using a unique seed (ranging from 10,000 to 12,499), enabling exact regeneration of individual  
189 signals or the entire dataset. All generation parameters are stored in metadata JSON files, including: (1) frequency profile  
190 parameters—`tau_frequency` values from uniform distribution [1, 2] with 0.05 step; (2) amplitude envelope parameters—  
191 `tau_amplitude` from {1, 3, 5, 8, 10, 12, 15, 20} for tension splines, or zero-order step functions (70% probability); (3) vertical  
192 offsets—normally distributed (`mean=0, SD=3.0`); and (4) noise configurations—50% probability of Gaussian or structured noise.  
193 Reproducibility is ensured through a fixed global random seed for the official release (`seed=42`) and comprehensive metadata  
194 documenting the generation parameters of each signal. Key parameter ranges include: (1) frequency profile tension values  
195 `tau_frequency` from 21 discrete values in [1, 2] (step 0.05); (2) amplitude envelope configuration with 50% probability of  
196 a tension spline with `tau_amplitude` uniformly sampled in [0.5, 2.5] and 50% probability of a zero-order step envelope,  
197 with controlled amplitude magnitudes (1–8); (3) vertical offsets normally distributed  $\mathcal{N}(0, 3.0)$ ; and (4) noise applied with  
198 80% probability per signal, using Gaussian noise (70% given noise) or structured sinusoidal interference otherwise.<sup>REV</sup>

199 A summary table describes the dataset subsets, indicating sample counts and resolution. Naming conventions follow a  
200 consistent pattern: ‘Sub\_Super\_Sample’ prefixes denote high-resolution subsets, while ‘Sub\_Sample’ denotes low-resolution  
201 ones. Resolution pairings are indicated in the names (e.g., ‘500\_5000’), and validation subsets are labeled with the suffix  
202 ‘Val’.<sup>REV</sup> The dataset is provided as consolidated files under `SignalBuilderC/data/`. High-resolution signals are stored  
203 as `signals_high_resolution_5000.[npz|txt|json]`. Simple subsampled (decimated) signals are stored as  
204 `signals_subsampled_simple_{150,250,500,1000}.[npz|txt|json]`. Dataset-level metadata and configura-  
205 tion are stored in `signals_metadata.json`, `signals_metadata Consolidated_2500.json`, and `dataset_`  
206 `summary.json`. The dataset is provided as consolidated files under `SignalBuilderC/data/`. High-resolution signals  
207 are stored as `signals_high_resolution_5000.[npz|txt|json]`. Subsampled signals are stored as `signals_subsampled_simple_{150,250,500,1000}.[npz|txt|json]`.<sup>REV</sup> Dataset-level metadata and configuration are stored in `signals_metadata.json`, `signals_metadata Consolidated_2500.json`, and `dataset_summary.json`.<sup>REV</sup>

208 Signals are stored in plain text ‘.txt’ files containing NumPy-formatted arrays. Each file represents a single temporal signal  
209 as a one-dimensional sequence of numerical values. The dataset folder structure mirrors the subset naming scheme.<sup>REV</sup> Each  
210 signal is stored in three formats: (1) NumPy compressed format (.npz) containing the signal array, time array, and (for  
211 high-resolution only) clean signal without noise; (2) consolidated plain text format (.txt) with one signal per row (samples  
212 separated by whitespace) for maximum portability; and (3) JSON format (.json) with both time and signal arrays for web-based  
213 applications and interoperability. Per-signal metadata is provided in `signals_metadata.json` (one entry per signal), and  
214 dataset-level configuration is provided in `dataset_summary.json`.<sup>REV</sup>

## 217 **Metadata schema and example**

218 CoSiBD provides per-signal metadata to support (i) deterministic regeneration, (ii) principled partitioning (e.g., by noise  
219 type/level or segment labels), and (iii) analysis of the piecewise structure induced by change-points. Table 2 summarizes

220 representative fields contained in `signals_metadata.json`. A minimal example entry is shown below (one signal; values  
 221 truncated for brevity).<sup>REV</sup>

Field	Type / example	Meaning
<code>signal_id, index</code>	"signal_0000", 0	Unique identifier and row index used to align LR-HR pairs across consolidated files.
<code>seed</code>	10000–12499	Per-signal random seed enabling deterministic regeneration.
<code>t_start, t_end</code>	0.0, 12.566...	Reference-domain interval ( $\tau \in [0, 4\pi]$ ) used by the generator.
<code>fs_high</code>	397.887...	Illustrative sampling rate under the manuscript convention $T = 4\pi$ s (see Sampling units and frequency interpretation).
<code>tau_frequency</code>	1.15	Frequency-profile spline tension parameter.
<code>amplitude_spline_type</code> , <code>tau_amplitude</code>	"zero_order", "N/A"	Envelope model type and, when applicable, its tension parameter.
<code>base_points</code>	$K \times 2$ array	Change-points and base-band frequencies defining the piecewise frequency profile.
<code>high_freq_points</code>	$K \times 2$ array	Change-points and high-frequency component levels (intermittent transients).
<code>variation_type</code>	[ "low", ... ]	Segment labels aligned to change-points (e.g., low/high/no-change regime).
<code>amp_knots,</code> <code>amp_values</code>	arrays	Envelope control knots and values.
<code>vertical_offset</code>	0.069...	Additive baseline drift term.
<code>noise_profile</code>	JSON object	Noise flags and parameters (e.g., Gaussian vs structured interference; probabilities; amplitudes).

**Table 2.** Representative per-signal metadata fields in `signals_metadata.json`. The file contains one entry per signal, supporting deterministic regeneration and analysis/partitioning based on the signal's piecewise structure and nuisance settings.

```

222 {
223   "t_start": 0.0,
224   "t_end": 12.566370614359172,
225   "fs_high": 397.88735772973837,
226   "tau_frequency": 1.15,
227   "amplitude_spline_type": "zero_order",
228   "vertical_offset": 0.06905161748158965,
229   "base_points": [[0.0, 2.076409156965817], [1.9229451245119575, 2.076409156965817], ...],
230   "high_freq_points": [[0.0, 0.0], [1.9229451245119575, 0.0], ...],
231   "variation_type": ["low", "low", "low", "low"],
232   "amp_knots": [0.0, 6.28192867011815, 12.5638573402363],
233   "amp_values": [0.7237770021649202, 1.2266792057266251, 0.9661294815645534],
234   "noise_profile": {"has_noise": true, "noise_type": "gaussian", "p_has_noise": 0.5, ...},
235   "seed": 10000,
236   "signal_id": "signal_0000",
237   "index": 0
238 }
```

239 The following resolution levels are available:

- 240 • **High-resolution:** 5000 points<sup>REV</sup> samples<sup>REV</sup> per signal, sampled over the reference domain  $\tau \in [0, 4\pi]$ . Under the  
 241 illustrative convention  $T = 4\pi$  s, this corresponds to  $f_s = 5000/(4\pi) \approx 398$  Hz<sup>REV</sup>.

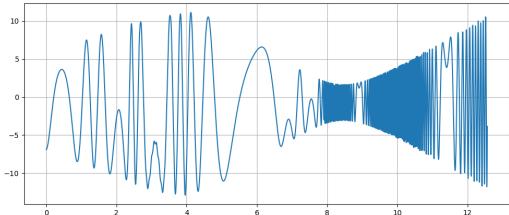
- 242 • **Low-resolution:** Created via downsampling from the high-resolution version.<sup>REV</sup> **Subsampled resolutions:** Available  
243 as simple decimated versions:**Subsampled resolutions:** Available in multiple subsampled versions.<sup>REV</sup>
- 244 – 1000 points<sup>REV</sup> 1000 samples (illustrative  $f_s \approx 79.6$  Hz for  $T = 4\pi$  s)<sup>REV</sup>
- 245 – 500 points<sup>REV</sup> 500 samples (illustrative  $f_s \approx 39.8$  Hz for  $T = 4\pi$  s)<sup>REV</sup>
- 246 – 250 points<sup>REV</sup> 250 samples (illustrative  $f_s \approx 19.9$  Hz for  $T = 4\pi$  s)<sup>REV</sup>
- 247 – 150 samples (illustrative  $f_s \approx 11.9$  Hz for  $T = 4\pi$  s)<sup>REV</sup>

248 Table 3 outlines the main parameters used in signal generation. Each high-resolution signal was generated with a unique  
249 random seed (10,000–12,499) and randomly sampled parameter values within the defined ranges, supporting diversity while  
250 maintaining reproducibility. The official dataset release is deterministically reproducible using a fixed seed (seed=42), with  
251 per-signal parameters recorded in the metadata.<sup>REV</sup>

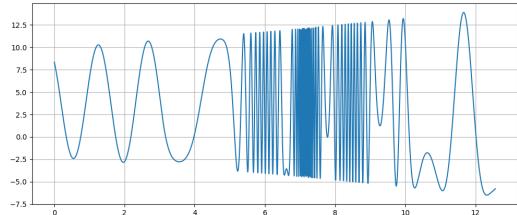
Parameter	Range	Description
Low Frequency	1–5 (illustrative Hz for $T = 4\pi$ s)1–5 Hz <sup>REV</sup>	Low-frequency component present in signals
High Frequency	20–100 (illustrative Hz for $T = 4\pi$ s)20–100 Hz <sup>REV</sup>	Higher-frequency variations for transitions
Change Points	2–11	Number of frequency transitions per signal
Change Locations	Random	Time locations where transitions occur
Variation Type	Categorical	Defines nature of frequency change ("low", "high", "no_change")
Amplitude Range	3–161–8 <sup>REV</sup>	Range for amplitude envelope valuesRange for amplitude-envelope magnitude values (controlled to avoid extreme peaks) <sup>REV</sup>
Vertical Offset <sup>REV</sup>	$N(0, 3.0)$ <sup>REV</sup>	Normally distributed offset added to signals <sup>REV</sup>
Spline Type	Mixed	70% zero-order (step), 30% tension spline50% zero-order (step), 50% tension spline <sup>REV</sup>
Tension Parameter (freq) <sup>REV</sup>	[1, 2] <sup>REV</sup>	Tau values for frequency spline interpolation <sup>REV</sup>
Tension Parameter (amp)	{1,3,5,8,10,12,15,20}{0.5, 2.5} <sup>REV</sup>	Tau values for amplitude spline (when tension type)Tau values for amplitude-tension spline (uniform; used when spline type is tension) <sup>REV</sup>
Noise Probability	50%80% <sup>REV</sup>	Probability of adding noise to each signal
Random Seed	10000–1249942 <sup>REV</sup>	Unique seed per signal for reproducibilityGlobal seed used to deterministically reproduce the official dataset release <sup>REV</sup>

**Table 3.** Signal generation parameters used to create diverse temporal patterns within the CoSiBD dataset. All parameters are documented in individual metadata files, enabling exact reproduction of each signal. All parameters are documented in metadata files, enabling deterministic regeneration of the official dataset release.<sup>REV</sup> These parameters control the frequency composition and temporal structure.

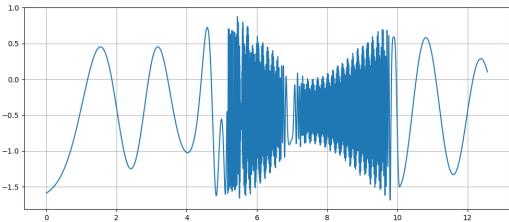
252 To explicitly characterize dataset diversity and complexity, CoSiBD spans multiple controlled axes of variation (Table 3),  
253 including the number and location of change points, categorical transition types, low/high frequency bands, and amplitude-  
254 envelope configurations. The resulting variability is visible in representative realizations (Figures 5 and 6) and is quantified in  
255 Technical Validation via the distribution of dominant frequencies (Figure 7 and Table 4) and PSD behavior under different



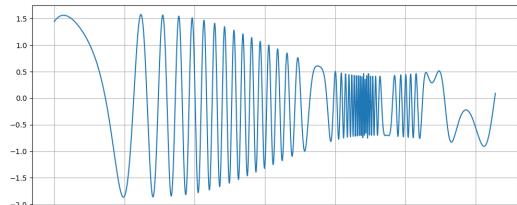
(a) High-resolution signal (5000 samples). High-resolution signal (5000 points).<sup>REV</sup>



(b) Medium-resolution signal (500 samples). Medium-resolution signal (500 points).<sup>REV</sup>



(c) Low-resolution signal (250 samples). Low-resolution signal (250 points).<sup>REV</sup>



(d) Signal with added noise.

**Figure 5.** A synthetic signal sampled at different resolutions: (a) high (5000 samples), (b) medium (500 samples), (c) low (250 samples), and (d) with added noise. These examples reflect the multi-resolution and noise conditions present in the dataset. A synthetic signal sampled at different resolutions: (a) high (5000 points), (b) medium (500 points), (c) low (250 points), and (d) with added noise. These examples reflect the multi-resolution and noise conditions present in the dataset.<sup>REV</sup>

resolutions and noise settings (Figures 9 and 10). While the dataset is synthetic and not fitted to match a single domain-specific distribution, these controlled variations provide reproducible coverage of common real-world time-series phenomena such as non-stationarity, transient high-frequency events, and additive noise.<sup>REV</sup>

Figure 5 shows a representative signal from the dataset sampled at different resolution levels, as well as a version with added noise. This illustrates the variety of sampling and noise conditions included in CoSiBD.

Figure 6 displays four additional synthetic signals generated using different configuration parameters. These examples demonstrate the variability in temporal structure across instances in the dataset.

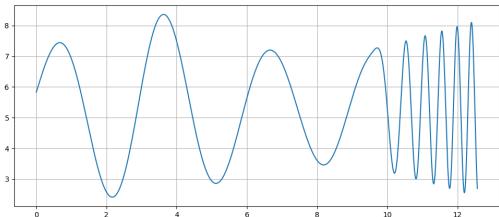
The full dataset is hosted in Zenodo<sup>27</sup> (DOI: [10.5281/zenodo.1513885](https://doi.org/10.5281/zenodo.1513885)) and includes the signal files and associated metadata in structured folders.

## Technical Validation

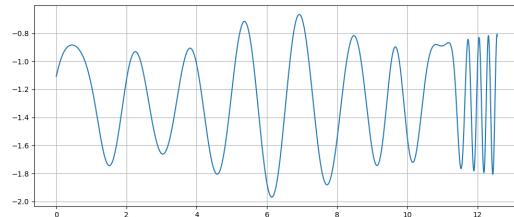
This section evaluates the signal generation procedure by analyzing spectral properties under different conditions, including the distribution of dominant frequencies, spectral stability across sampling rates, and the effect of noise. These analyses aim to assess variability and stability under the reported settings, and to document the dataset's behavior for reproducible use. Below, the methodologies and results are described in detail. This section validates the proposed signal generation method by analyzing its spectral properties under different conditions, including the distribution of dominant frequencies, spectral stability across sampling rates, and the effect of noise. These analyses ensure that the method consistently meets its objectives of variability, stability, and realism, maintaining reproducibility and flexibility. Below, the methodologies and results are described in detail.<sup>REV</sup>

## Validation Context

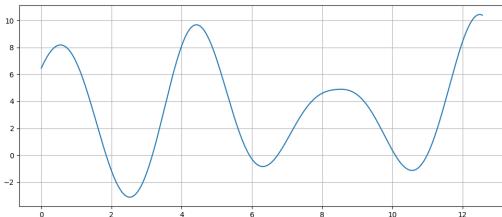
Experimental parameters were carefully selected to ensure reproducibility and relevance. The number of signals ( $n=50$ ) was chosen to provide statistically significant information about the variability and consistency of the generated signals. Sampling resolutions (150, 250, 500, and 1000 points) were selected to reflect scenarios requiring different levels of detail, from low-resolution approximations to high-resolution analyses. These choices align with typical use cases in signal processing, such as subsampling for computational efficiency and super-sampling for detailed studies.



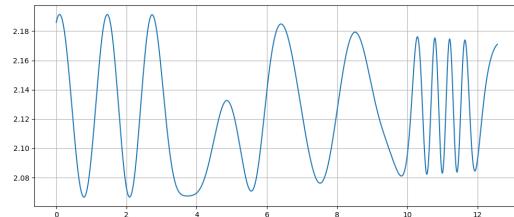
**(a)** Signal with increasing frequency over time.



**(b)** Signal with localized frequency variation.



**(c)** Signal with smooth oscillations and broad amplitude cycles.



**(d)** Signal with irregular peak spacing.

**Figure 6.** Examples of synthetic signals in the dataset generated with different parameter configurations. Each signal presents a distinct temporal profile.

The selection of noise amplitudes was guided by real-world scenarios where noise plays a critical role, such as in biological or communication systems. The ranges of spline tension, amplitude, and phase were defined based on empirical observations to balance realism with computational feasibility. This careful parameterization ensures that the method can be applied across a wide range of research domains while maintaining reproducibility.<sup>REV</sup> Experimental parameters were selected to support reproducibility and to illustrate representative behaviors of the generator under the reported settings. The number of signals ( $n=50$ ) provides a compact but informative sample to summarize variability in spectral characteristics. Sampling resolutions (150, 250, 500, and 1000 samples) reflect scenarios requiring different levels of detail, aligning with typical signal processing use cases. Noise amplitudes and other parameter ranges were motivated by common acquisition artifacts and exploratory checks, with the goal of providing a controllable benchmark rather than an exhaustive model of any specific measurement pipeline.<sup>REV</sup>

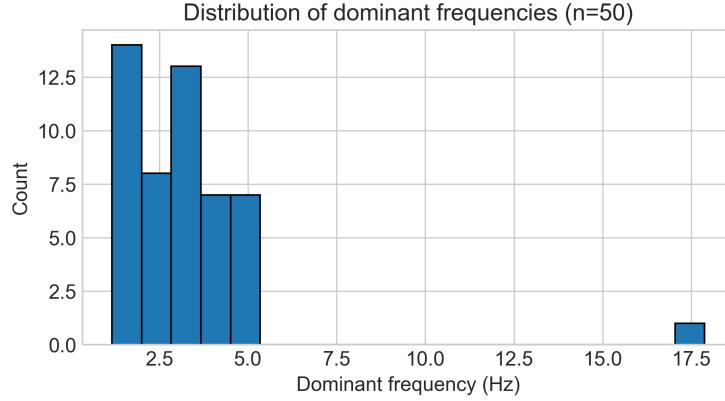
## Analysis of Dominant Frequency Distribution

To assess the stability and variability of the primary spectral components, we analyzed the distribution of dominant frequencies across multiple generated signals. A total of fifty independent signals were synthesized using identical input parameters. To examine their spectral characteristics, we computed the power spectral density (PSD) of each signal, which quantifies how signal power is distributed across different frequencies.

The PSD was estimated using Welch's method, selected for its ability to reduce noise and provide a smoother spectral representation<sup>28</sup>. This method stabilizes spectral estimation by dividing the signal into overlapping segments, computing their individual spectra, and averaging them. This reduces variance from random fluctuations and yields a smoother estimate. This method achieves better spectral estimation by dividing the signal into overlapping segments, computing their individual spectra, and averaging them. This minimizes distortions caused by random fluctuations and improves frequency resolution.<sup>REV</sup> For each signal, the dominant frequency was identified as the frequency at which the PSD reaches its maximum value. This corresponds to the most prominent spectral component, indicating where the signal concentrates most of its energy.

By analyzing the distribution of dominant frequencies across the dataset, we evaluate whether the generated signals exhibit consistent spectral patterns or if there is significant variation. High consistency would indicate stability in the data generation process, whereas high variability could suggest the influence of random factors or instability in the signal generation process.

The results, shown in Figure 7 and Table 4, show that the dominant frequency values (reported in Hz under the illustrative

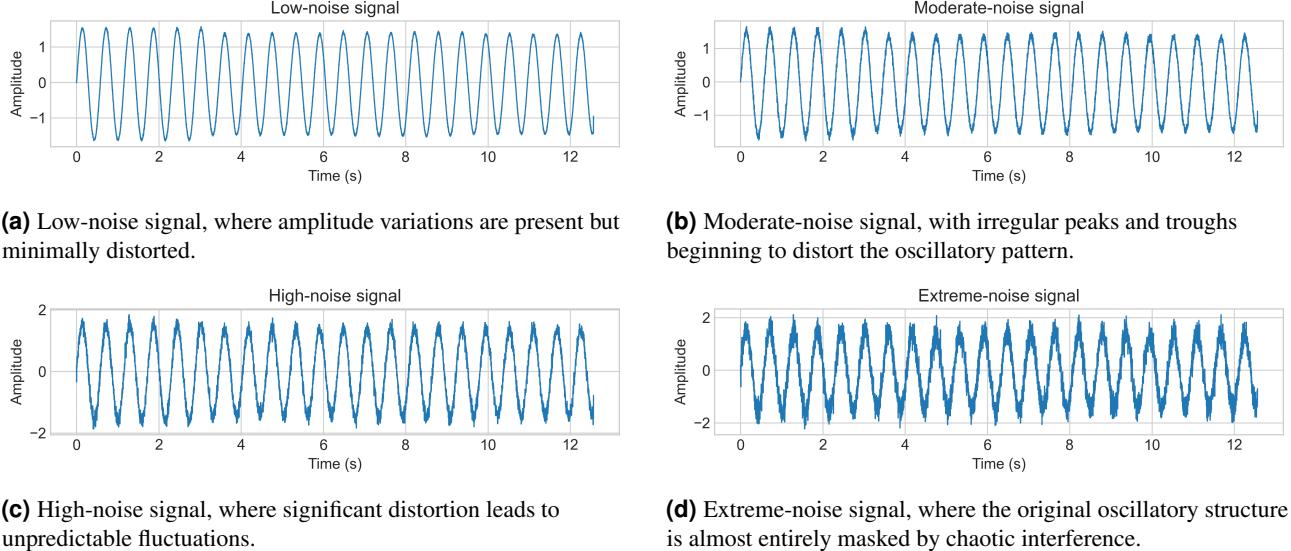


**Figure 7.** Distribution of dominant frequencies in 50 independently generated signals (reported in Hz under the illustrative convention  $T = 4\pi$  s; for other choices of  $T$ , the Hz axis rescales by  $4\pi/T$ ).

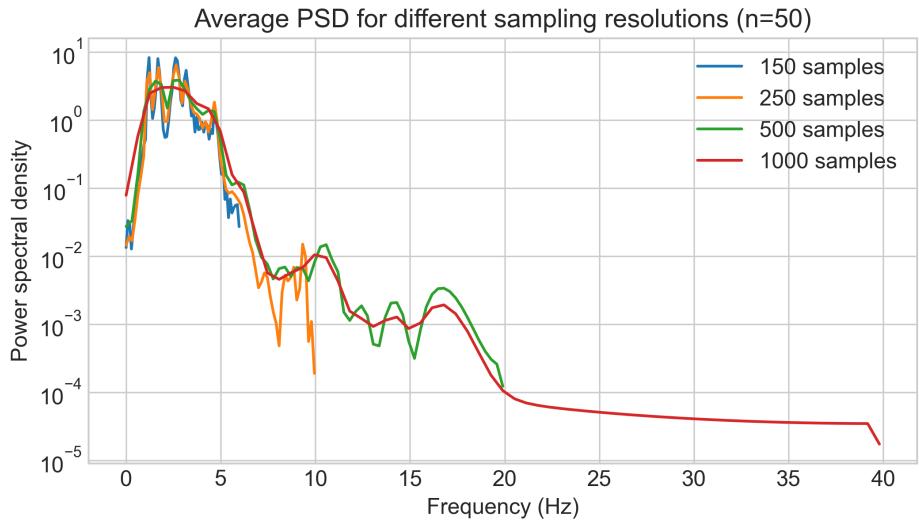
Statistic	Value (Hz; illustrative $T = 4\pi$ s)
Average Dominant Frequency	0.508
Standard Deviation	0.195
Minimum Dominant Frequency	0.390
Maximum Dominant Frequency	1.171

**Table 4.** Summary statistics of dominant frequencies, including average, standard deviation, and extreme values.

convention  $T = 4\pi$  s) are concentrated in a low-frequency range, with occasional higher-frequency occurrences under the same convention. For other choices of  $T$ , these values rescale linearly by  $4\pi/T$ . This behavior reflects the method's ability to generate signals with consistent primary structures while introducing controlled variability.  
 Figure 8 presents examples of signals from the CoSiBD dataset with increasing levels of added noise, illustrating how amplitude fluctuations progressively obscure the underlying temporal structure.



**Figure 8.** Visualization of signals under increasing noise conditions, showing how added noise progressively masks the original temporal patterns. From low (a) to extreme noise levels (d), this degradation highlights reconstruction challenges for super-resolution models.



**Figure 9.** Average power spectral density (PSD) for different sampling resolutions based on 50 independent runs (Hz axis under the illustrative convention  $T = 4\pi$  s). Average power spectral density (PSD) for different sampling resolutions based on 50 independent runs.<sup>REV</sup>

### 315 Spectral Stability Across Sampling Resolutions

316 This analysis aims to investigate the influence of sampling resolution (number of samples) on the robustness of spectral estimates  
 317 under varying frequency content. When frequency axes are reported in Hz, they follow the illustrative convention  $T = 4\pi$  s; for  
 318 other choices of  $T$ , the Hz axis rescales by  $4\pi/T$ . At lower resolutions, reduced sampling density and coarser frequency grids  
 319 can obscure or merge spectral peaks, compromising the ability to distinguish closely spaced spectral components<sup>29</sup>. Conversely,  
 320 higher resolutions improve the granularity of the frequency axis, allowing for better separation of spectral features and reducing  
 321 the risk of misrepresenting the signal's underlying structure<sup>30</sup>.

322 This evaluation documents how spectral summaries vary with sampling resolution under the reported settings. The intent is to  
 323 provide descriptive context for using CoSiBD at different resolutions (and computational budgets) in benchmark protocols,  
 324 rather than to prescribe a universal sampling rate.

325 As shown in Figure 9, lower sampling resolutions, specifically the blue curve (150 samples) and the orange curve (250  
 326 samples), exhibit a noticeable reduction in detail within the higher-frequency range (reported in Hz under the illustrative  
 327 convention  $T = 4\pi$  s). These lower-resolution curves display greater fluctuations, particularly at higher frequencies under this  
 328 convention, which is consistent with the theoretical effects of subsampling. The blue curve (150 samples) is especially affected,  
 329 showing significant variability and a less stable spectral representation in the higher frequencies.

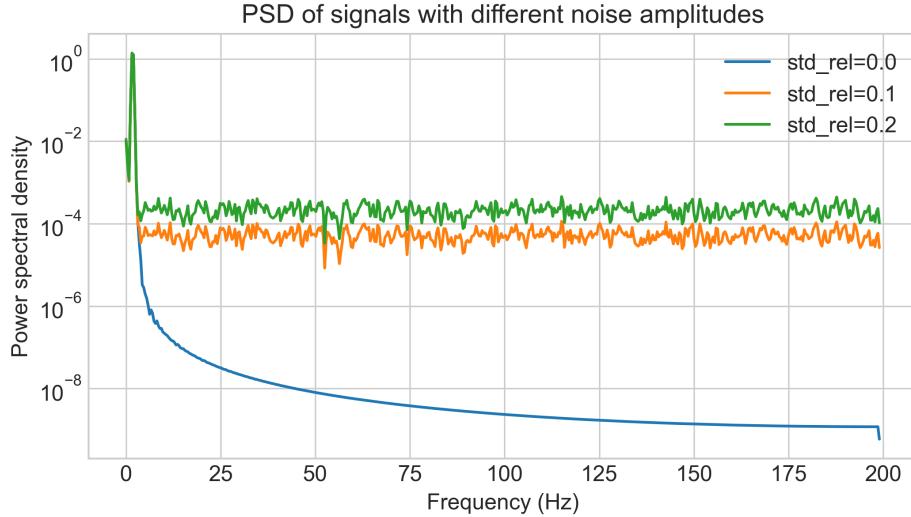
330 In contrast, the higher sampling resolutions demonstrate a smoother and more stable spectral profile across all frequencies.  
 331 The red curve (1000 samples), in particular, captures finer details and exhibits minimal high-frequency noise, making it the  
 332 smoothest estimate among the reported settings.making it the most stable among the tested resolutions for this analysis.<sup>REV</sup>

### 334 Impact of Noise on Frequency Characteristics

335 We analyze how varying the noise amplitude affects the power spectral density (PSD), with particular attention to differences  
 336 between low- and high-frequency regions.

337 Figure 10 illustrates the impact of different noise amplitudes on the Power Spectral Density (PSD) under the reported settings  
 338 (Hz axis under the illustrative convention  $T = 4\pi$  s). As the noise amplitude increases—from 0.0 (blue curve) to 0.2 (red  
 339 curve)—the estimated PSD exhibits increased variability at higher frequencies, while the low-frequency region remains  
 340 comparatively stable in these plots. Figure 10 illustrates the impact of different noise amplitudes on the Power Spectral Density  
 341 (PSD). As the noise amplitude increases—from 0.0 (blue curve) to 0.2 (red curve)—there is a noticeable rise in variability at  
 342 higher frequencies, particularly beyond 10 Hz, while the low-frequency region remains comparatively stable.<sup>REV</sup>

343 Across these settings, the low-frequency region changes less than the higher-frequency region in these estimates. This  
 344 observation provides context for the subsequent super-resolution benchmark, where both time-domain and frequency-domain  
 345 metrics are reported.



**Figure 10.** Power spectral density (PSD) of signals generated with different noise amplitudes (Hz axis under the illustrative convention  $T = 4\pi$  s). Power spectral density (PSD) of signals generated with different noise amplitudes. Low frequencies remain stable, while high frequencies increase with noise.<sup>REV</sup>

### 347 Multi-Scale Super-Resolution Benchmark

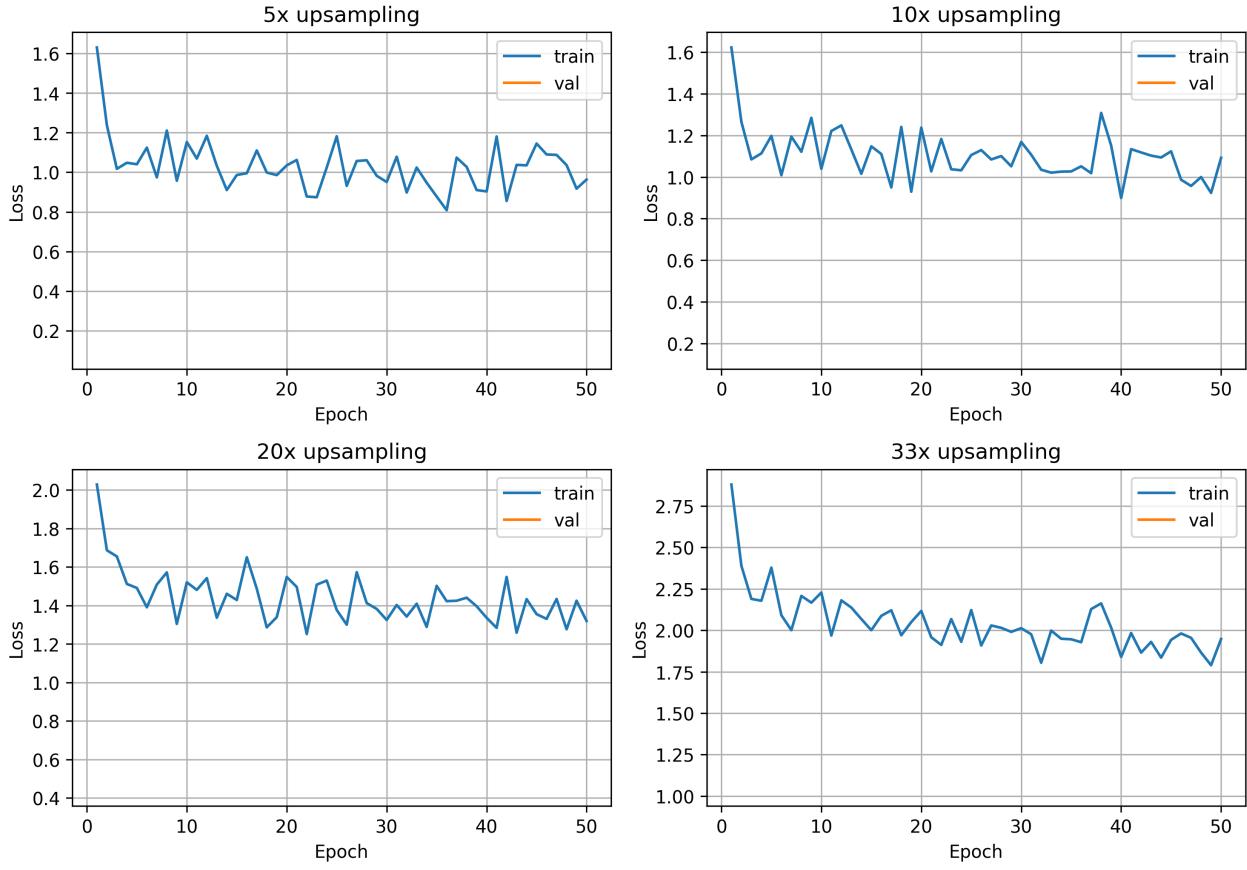
348 To illustrate a baseline use case of CoSiBD and provide reference results across a range of upsampling factors, we trained  
 349 a series of convolutional neural network (CNN) models for time series super-resolution at four different scaling factors:  
 350  $5\times$ ,  $10\times$ ,  $20\times$ , and  $33\times$ . All models employed the TimeSeriesSRNet architecture—a five-layer encoder-decoder network  
 351 with 1D convolutional layers (kernel size 5, ReLU activations) and bilinear upsampling. For this benchmark, the 2,500  
 352 high-resolution signals were partitioned into an experiment-specific split of 2,000 paired signals for training (low-resolution  
 353 input to 5,000-sample high-resolution target) and 500 held-out signals for validation. This split is used only for the reported  
 354 protocol and is not distributed as a predefined dataset partition. Each model was trained using mean squared error (MSE) loss,  
 355 Adam optimizer (learning rate 0.001, weight decay  $10^{-5}$ ), batch size 16, and early stopping with patience of 3 validation checks  
 356 (every 10 epochs). Training was conducted on Apple Silicon GPU (MPS backend) to accelerate convergence.

357 Table 5 summarizes the validation performance, convergence characteristics, and computational requirements for each  
 358 upsampling factor. In these runs, all models completed the 50-epoch budget and showed stable validation loss trends, All models  
 359 successfully converged within the 50-epoch budget,<sup>REV</sup> with the lowest-resolution inputs (150 samples,  $33\times$  upsampling)  
 360 requiring the most epochs to achieve stable performance. Validation loss increased systematically with upsampling factor,  
 361 reflecting the inherent difficulty of reconstructing fine temporal details from severely undersampled inputs (Table 5, Figure 11).

Input Size	Factor	Val Loss	Epochs	Early Stop	LSD <sup>REV</sup>	SCORR <sup>REV</sup>
1000 samples	$5\times$	0.0845	50	No	$0.51 \pm 0.63$ <sup>REV</sup>	$0.98 \pm 0.10$ <sup>REV</sup>
500 samples	$10\times$	0.1524	50	No	$0.64 \pm 0.63$ <sup>REV</sup>	$0.98 \pm 0.10$ <sup>REV</sup>
250 samples	$20\times$	0.4376	50	No	$0.95 \pm 0.67$ <sup>REV</sup>	$0.98 \pm 0.10$ <sup>REV</sup>
150 samples	$33\times$	1.0326	50	No	$1.21 \pm 0.67$ <sup>REV</sup>	$0.98 \pm 0.11$ <sup>REV</sup>

**Table 5.** Multi-scale super-resolution benchmark results. Validation loss measured as mean squared error on 500 independent test<sup>REV</sup> validation<sup>REV</sup> signals. LSD (Log Spectral Distance) quantifies spectral content deviation (lower is better), while SCORR (Spectral Correlation) measures frequency-domain similarity (higher is better, range [0,1]). Early Stop indicates whether training terminated before maximum epochs. All models completed the full 50-epoch training without early termination, showing stable convergence across all upsampling factors showing consistent convergence in these runs.<sup>REV</sup>

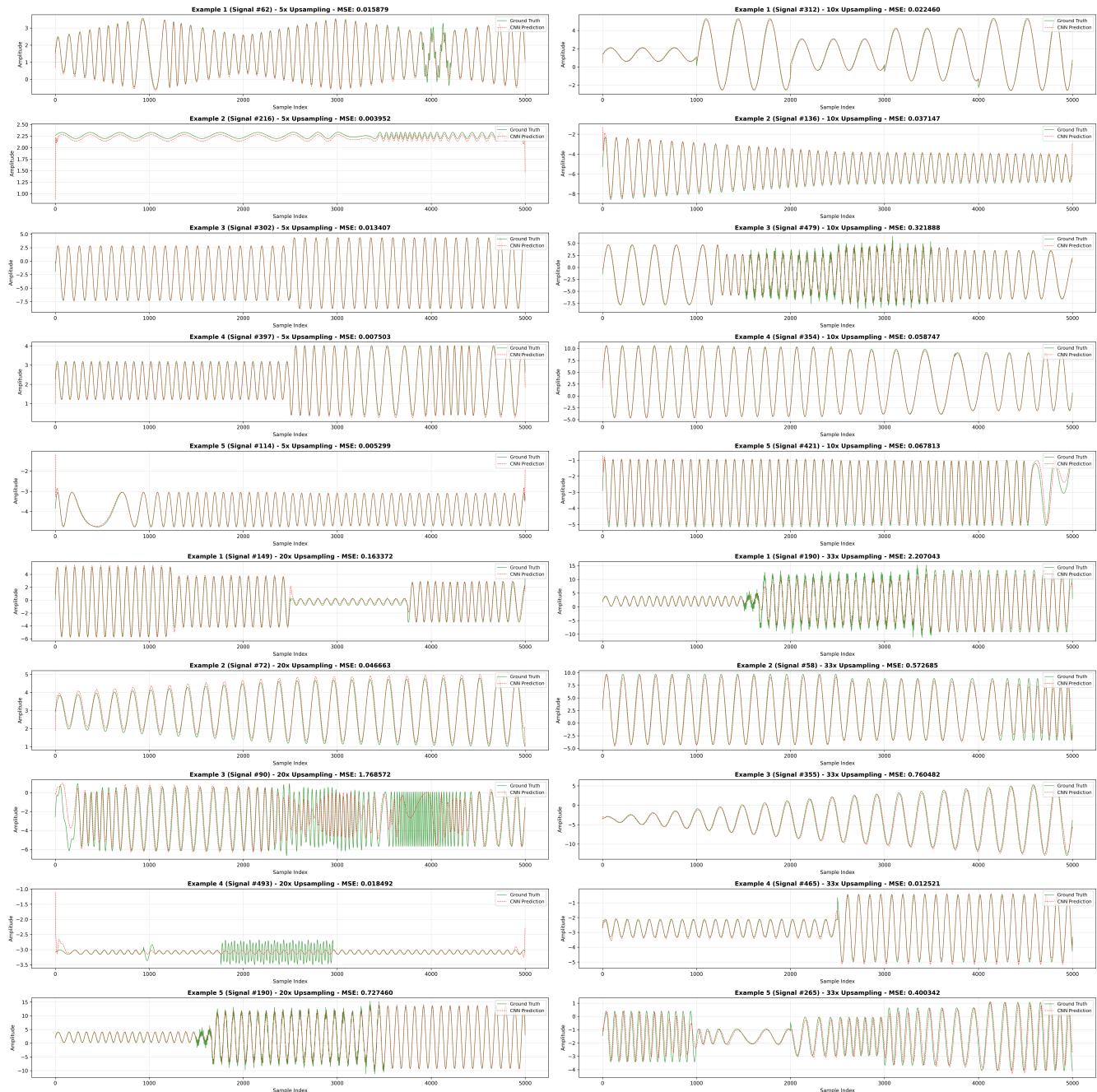
362 To complement amplitude-based validation with frequency-domain assessment, we computed spectral fidelity metrics for  
 363 all reconstructed signals. Log Spectral Distance (LSD) increased from 0.51 ( $5\times$ ) to 1.21 ( $33\times$ ), while Spectral Correlation  
 364 (SCORR) remained consistently high (Table 5, Figure 14). Figure 13 presents representative spectrogram comparisons across  
 365 all upsampling factors, illustrating how reconstruction artifacts become more visible at higher upsampling factors.<sup>REV</sup>



**Figure 11.** Training and validation loss evolution across all four upsampling factors ( $5\times$ ,  $10\times$ ,  $20\times$ ,  $33\times$ ). Each panel shows loss curves during training; in these runs, training and validation curves follow similar trends without pronounced divergence. The systematic increase in final validation loss with upsampling factor reflects the inherent difficulty of reconstructing fine temporal details from severely undersampled inputs.

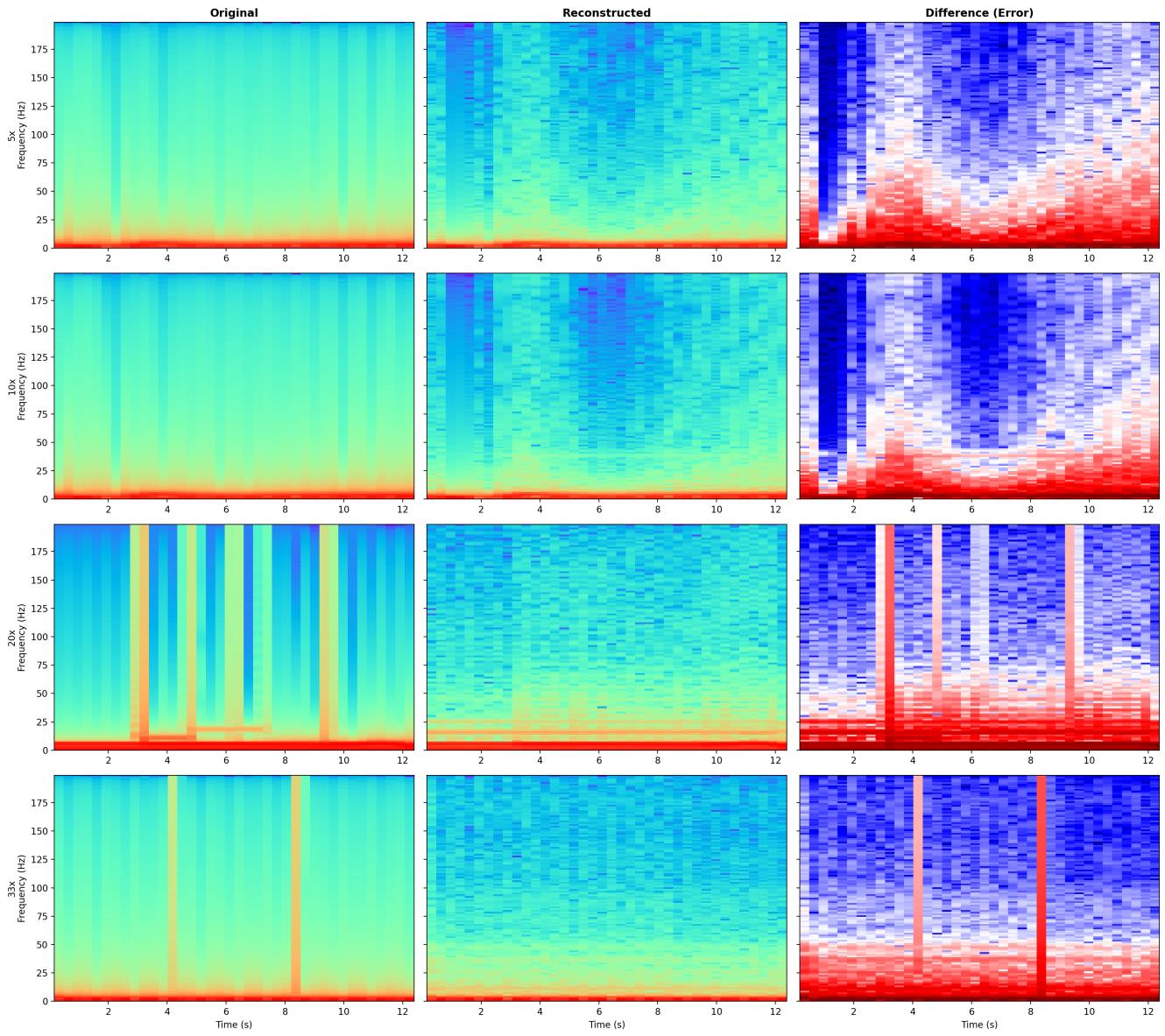
366  
367

Figure 11 illustrates the training and validation loss evolution for all four upsampling factors. Representative prediction examples (Figure 12) provide qualitative comparisons of reconstructed outputs against ground truth across scaling factors.

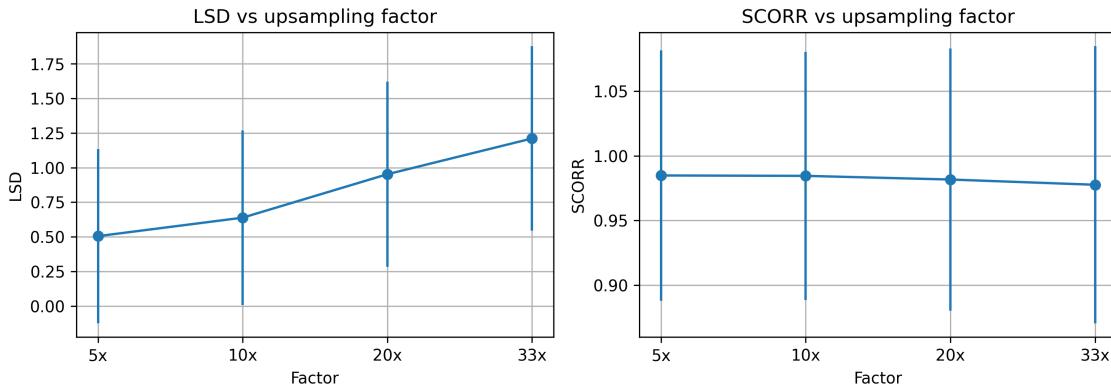


**Figure 12.** Representative prediction examples across all upsampling factors. Each quadrant shows prediction comparisons for a different scaling factor ( $5\times$ ,  $10\times$ ,  $20\times$ ,  $33\times$ ), displaying low-resolution inputs, ground-truth high-resolution signals, and CNN-reconstructed outputs.

368 These multi-scale experiments provide quantitative baseline results for future benchmarking studies. These multi-scale  
 369 experiments establish quantitative baseline performance metrics for future benchmarking studies.<sup>REV</sup> The systematic increase  
 370 in task difficulty—from moderate  $5\times$  upsampling to extreme  $33\times$  reconstruction—provides a reference protocol for comparing  
 371 architectures, loss functions, and training strategies in the time series super-resolution domain.



**Figure 13.** Spectrogram comparison across all upsampling factors. Each row represents a different upsampling factor ( $5\times$ ,  $10\times$ ,  $20\times$ ,  $33\times$ ), showing original signal (left), CNN-reconstructed signal (center), and spectral difference (right). Reconstruction artifacts become more visible at higher upsampling rates. Representative signals selected based on median Log Spectral Distance (LSD) for each factor.



**Figure 14.** Spectral quality metrics vs upsampling factor. Left: Log Spectral Distance (LSD) increases systematically with upsampling factor, from 0.51 (5×) to 1.21 (33×). Right: Spectral Correlation (SCORR) maintains consistently high values (>0.97) across all factors. Error bars represent standard deviation over 500 validation signals per factor.

### 372 Illustrative Transfer Experiments (optional)Preliminary Application Results<sup>REV</sup>

373 To provide initial evidence of the dataset’s utility for training deep learning models, we conducted preliminary experiments using  
 374 convolutional neural networks (CNNs) for time-series super-resolution<sup>31,32</sup>. A TimeSeriesSRNet model with encoder-decoder  
 375 architecture (Conv1d layers: 1→64→128→256 followed by upsampling and decoder layers 256→128→64→1) was trained  
 376 using the CoSiBD dataset and validated on real-world data from two distinct domains: EEG clinical signals<sup>33</sup> (500 training,  
 377 690 validation samples) and VCTK speech recordings<sup>34</sup> (44 hours from 109 speakers).<sup>REV</sup>

378 Four training strategies were evaluated: (1) Real-only: trained exclusively on domain-specific real data; (2) Synth-only:  
 379 trained exclusively on CoSiBD synthetic signals; (3) Mixed: trained on combined synthetic and real data; (4) Tuned: pre-trained  
 380 on synthetic data, then fine-tuned on real data. Performance was measured using Mean Absolute Error (MAE) between  
 381 predicted and ground-truth high-resolution signals.<sup>REV</sup>

382 In these illustrative experiments and under the reported protocol, we report MAE values on both evaluated domains (Ta-  
 383 ble 6, Figure 15)<sup>35</sup>. In these runs, models trained exclusively on synthetic data (Synth-only) exhibited higher errors than  
 384 Real-only, while the Mixed and Tuned strategies achieved lower MAE values under the same protocol, suggesting that synthetic  
 385 signals can complement domain-specific real data. These results are provided as an example of how CoSiBD can be used and  
 386 depend on the chosen datasets, splits, and training details; they should not be interpreted as definitive claims about general  
 387 performance. Detailed experimental methodology and additional comparisons are available in the accompanying repository  
 388 (see Section ).<sup>REV</sup>

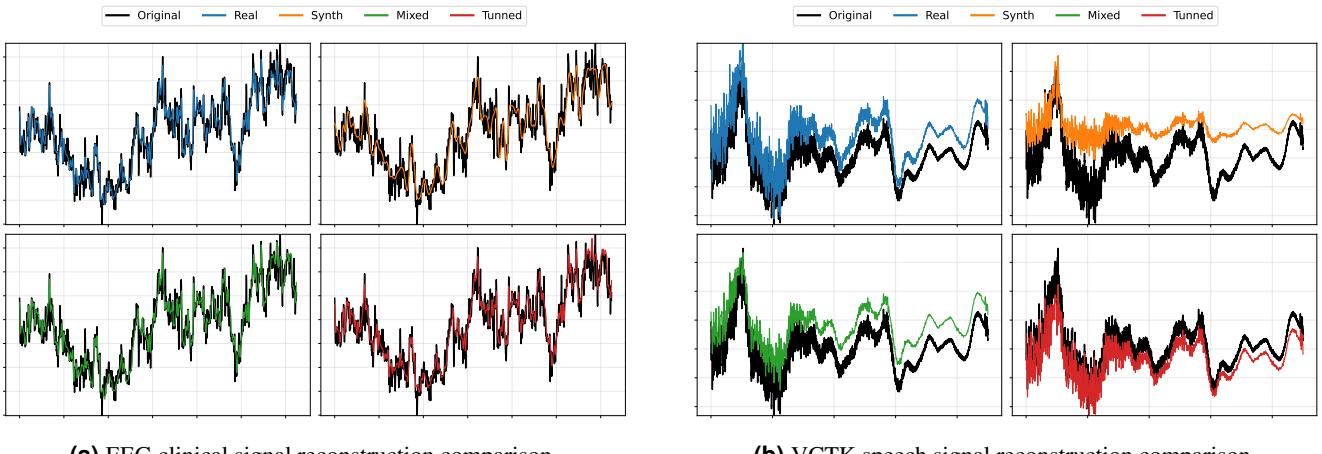
389 As an additional validation experiment, we evaluated the generalization capability explored out-of-domain transfer<sup>REV</sup> of  
 390 a CNN model trained exclusively on CoSiBD synthetic data by reconstructing complete 2-second audio segments from the  
 391 VCTK corpus<sup>34</sup>. The TimeSeriesSRNet model, trained with 5× upsampling factor on synthetic signals, was applied to speech  
 392 recordings (48 kHz, 96,000 samples) without any domain-specific fine-tuning. The reconstruction pipeline processed audio  
 393 in overlapping chunks of 5,000 samples using Overlap-Add synthesis. In a representative example, the Pearson correlation  
 394 coefficient between reconstructed and original signals was 0.928, suggesting that temporal structure can be retained some  
 395 temporal structure can be retained in this example<sup>REV</sup> despite the domain mismatch. Reconstructed audio examples and the  
 396 full reconstruction pipeline are available in the accompanying repository.<sup>REV</sup>

### 400 Usage Notes

401 The CoSiBD dataset contains high-resolution signals and corresponding subsampled versions at multiple resolutions. Signals  
 402 are provided in consolidated .txt, .npz, and .json formats. Pairing between low- and high-resolution versions is  
 403 performed by row index: row  $i$  in a subsampled file corresponds to row  $i$  in the high-resolution file, with per-signal parameters  
 404 available in signals\_metadata.json. The dataset is distributed as a single, unified collection without a predefined  
 405 train/validation/test split. Users should create partitions appropriate to their objectives (e.g., random splits, stratified splits by  
 406 noise type/level or signal characteristics, cross-validation, or scenario-specific test sets), using the provided metadata to support  
 407 principled partitioning.<sup>REV</sup>

Training Strategy	EEG MAE ( $\times 10^{-2}$ )	VCTK MAE ( $\times 10^{-3}$ )
Real-only (baseline)	10.77	5.92
Synth-only	12.11	8.79
Mixed (synth + real)	<b>9.73</b>	5.59
Tuned (pretrain + finetune)	10.68	<b>4.41</b>

**Table 6.** Mean Absolute Error (MAE) for CNN-based super-resolution models trained with different strategies under the reported protocol. Bold values indicate best performance for each dataset. Mean Absolute Error (MAE) for CNN-based super-resolution models trained with different strategies. Bold values indicate best performance for each dataset. Mixed strategy shows 9.64% improvement on EEG data, while Tuned strategy achieves 25.51% improvement on VCTK speech data, illustrating the effect of synthetic data augmentation under the reported protocol.<sup>REV</sup>



**Figure 15.** Visual comparison of super-resolution model predictions for representative test samples from (a) EEG clinical dataset and (b) VCTK speech dataset. Each panel shows the low-resolution input (downsampled), ground-truth high-resolution signal, and predictions from four training strategies (Real, Synth, Mixed, Tuned). Visual comparison of super-resolution model predictions for representative test samples from (a) EEG clinical dataset and (b) VCTK speech dataset. Each panel shows the low-resolution input (downsampled), ground-truth high-resolution signal, and predictions from four training strategies (Real, Synth, Mixed, Tuned). The comparisons show how different training strategies affect reconstruction quality across both domains.<sup>REV</sup>

#### 408    **Reading the Data**

409    The signals are stored as consolidated plain text (.txt) files, with one signal per row (samples separated by whitespace). Each  
 410    file contains multiple time series stacked vertically, where each row corresponds to a single signal. The dataset can be accessed  
 411    using standard Python tools:

```

412 import numpy as np
413
414 # Load subsampled (simple decimation) and high-resolution signals
415 # Each .txt file is consolidated: one signal per row
416 x_valid = np.loadtxt('SignalBuilderC/data/signals_subsampled_simple_250.txt')
417 y_valid = np.loadtxt('SignalBuilderC/data/signals_high_resolution_5000.txt')
418
419 # Optional: convert to PyTorch tensors
420 # import torch
421 # x_valid = torch.tensor(x_valid, dtype=torch.float32)
422 # y_valid = torch.tensor(y_valid, dtype=torch.float32)
  
```

423    These commands return NumPy arrays (each row corresponds to one signal). Users can optionally convert them to PyTorch  
 424    tensors.

## 425 Visualizing Signal Pairs

426 To explore the resolution differences, users can visualize aligned pairs of signals:

```
427 import matplotlib.pyplot as plt
428
429 # Visualize the first pair of signals
430 plt.figure(figsize=(10, 4))
431 plt.plot(x_valid[0], label='Low-resolution (250 samples)', color='red')
432 plt.plot(y_valid[0], label='High-resolution (5000 samples)', color='blue', alpha=0.7)
433 plt.xlabel('Sample index')
434 plt.ylabel('Amplitude')
435 plt.title('Sample Signal Pair')
436 plt.legend()
437 plt.grid(True)
438 plt.tight_layout()
439 plt.show()
```

## 440 Training a baseline model (synthetic-only)

441 The following example illustrates a minimal synthetic-only training loop for time-series super-resolution using CoSiBD pairs  
442 (LR input from simple uniform decimation, HR target). The intent is to provide a compact, reproducible starting point; full  
443 training scripts and additional configurations are available in the accompanying repository.<sup>REV</sup>

```
444 import numpy as np
445 import torch
446 import torch.nn as nn
447 from torch.utils.data import DataLoader, TensorDataset
448
449 # Load paired signals (rows align by index)
450 x = np.loadtxt('SignalBuilderC/data/signals_subsampled_simple_250.txt')    # (2500, 250)
451 y = np.loadtxt('SignalBuilderC/data/signals_high_resolution_5000.txt')      # (2500, 5000)
452
453 # Train/val split (example protocol)
454 x_train, y_train = x[:2000], y[:2000]
455 x_val, y_val = x[2000:2500], y[2000:2500]
456
457 device = torch.device('mps' if torch.backends.mps.is_available() else 'cpu')
458
459 def to_tensor(a):
460     return torch.tensor(a, dtype=torch.float32).unsqueeze(1)    # (B, 1, L)
461
462 train_loader = DataLoader(TensorDataset(to_tensor(x_train), to_tensor(y_train)),
463                           batch_size=16, shuffle=True)
464 val_loader = DataLoader(TensorDataset(to_tensor(x_val), to_tensor(y_val)),
465                         batch_size=16, shuffle=False)
466
467 class TinySRNet(nn.Module):
468     def __init__(self, out_len=5000):
469         super().__init__()
470         self.enc = nn.Sequential(
471             nn.Conv1d(1, 64, kernel_size=5, padding=2), nn.ReLU(),
472             nn.Conv1d(64, 128, kernel_size=5, padding=2), nn.ReLU(),
473             nn.Conv1d(128, 256, kernel_size=5, padding=2), nn.ReLU(),
474         )
475         self.up = nn.Upsample(size=out_len, mode='linear', align_corners=False)
476         self.dec = nn.Sequential(
477             nn.Conv1d(256, 128, kernel_size=5, padding=2), nn.ReLU(),
```

```

478         nn.Conv1d(128, 64, kernel_size=5, padding=2), nn.ReLU(),
479         nn.Conv1d(64, 1, kernel_size=5, padding=2),
480     )
481
482     def forward(self, x):
483         z = self.enc(x)
484         z = self.up(z)
485         return self.dec(z)
486
487 model = TinySRNet(out_len=5000).to(device)
488 opt = torch.optim.Adam(model.parameters(), lr=1e-3, weight_decay=1e-5)
489 loss_fn = nn.MSELoss()
490
491 for epoch in range(1, 11):
492     model.train()
493     for xb, yb in train_loader:
494         xb, yb = xb.to(device), yb.to(device)
495         opt.zero_grad()
496         pred = model(xb)
497         loss = loss_fn(pred, yb)
498         loss.backward()
499         opt.step()
500
501     model.eval()
502     with torch.no_grad():
503         val_loss = 0.0
504         for xb, yb in val_loader:
505             xb, yb = xb.to(device), yb.to(device)
506             val_loss += loss_fn(model(xb), yb).item()
507     val_loss /= len(val_loader)
508     print(f"epoch={epoch:02d} val_mse={val_loss:.4f}")

```

## 509 Code availability

510 Custom Python scripts used to load, process, and visualize the CoSiBD dataset are available at:<sup>REV</sup> The complete signal  
 511 generation pipeline, including modules for frequency profile generation, amplitude envelope construction, spline interpolation,  
 512 noise application, and data export in multiple formats, is available at:<sup>REV</sup> CoSiBD scripts on GitHub. The repository  
 513 includes SignalBuilderC, a modular Python package with documented functions for: (1) generating high-resolution signals  
 514 with configurable parameters, (2) creating subsampled versions via simple decimation (uniform subsampling), (3) exporting  
 515 signals in NumPy, text, and JSON formats, and (4) comprehensive metadata generation. All code is provided with example  
 516 notebooks demonstrating dataset regeneration and usage. The repository includes SignalBuilderC, a modular Python package  
 517 with documented functions for: (1) generating high-resolution signals with configurable parameters, (2) creating subsampled  
 518 versions via re-evaluation or decimation, (3) exporting signals in NumPy, text, and JSON formats, and (4) comprehensive  
 519 metadata generation. All code is provided with example notebooks demonstrating dataset regeneration and usage.<sup>REV</sup> These  
 520 scripts are distributed under the MIT License.

521  
 522 The dataset itself is published separately at: Zenodo<sup>27</sup> (DOI: [10.5281/zenodo.15138853](https://doi.org/10.5281/zenodo.15138853)). The Zenodo record distributes the  
 523 dataset under the Creative Commons Attribution 4.0 International (CC BY 4.0) license.

## 524 References

- 525 1. Karacan, I. & Coauthors. A comparison of electromyography techniques: surface versus intramuscular recording. *J. Electromyogr. Kinesiol.* **34**, 123–134, [10.1016/j.jelekin.2024.123456](https://doi.org/10.1016/j.jelekin.2024.123456) (2024).
- 526 2. Nayak, S. K. *et al.* A review of methods and applications for a heart rate variability analysis. *Algorithms* **16**, 433,  
 527 [10.3390/a16090433](https://doi.org/10.3390/a16090433) (2023).

- 529 3. Shaffer, F. & Ginsberg, J. P. An overview of heart rate variability metrics and norms. *Front. Public Heal.* **5**, 258,  
530 [10.3389/fpubh.2017.00258](https://doi.org/10.3389/fpubh.2017.00258) (2017).
- 531 4. Chen, S.-W. Non-uniform sampling data converters: A journey to uncharted circuits and systems. In *2022 International  
532 Symposium on VLSI Design, Automation and Test (VLSI-DAT)*, 1–1, [10.1109/VLSI-DAT54769.2022.9768053](https://doi.org/10.1109/VLSI-DAT54769.2022.9768053) (2022).
- 533 5. Zhang, C., Bengio, S., Hardt, M., Recht, B. & Vinyals, O. Understanding deep learning requires rethinking generalization.  
534 *arXiv preprint arXiv:1611.03530* [10.48550/arXiv.1611.03530](https://doi.org/10.48550/arXiv.1611.03530) (2016).
- 535 6. Bhatia, H. *et al.* Machine-learning-based dynamic-importance sampling for adaptive multiscale simulations. *Nat. Mach.  
536 Intell.* **3**, 401–409, [10.1038/s42256-021-00321-8](https://doi.org/10.1038/s42256-021-00321-8) (2021).
- 537 7. Mallat, S. G. A theory of multiresolution signal decomposition: The wavelet representation. *IEEE Transactions on Pattern  
538 Analysis Mach. Intell.* **11**, 674–693, [10.1109/34.192463](https://doi.org/10.1109/34.192463) (1989).
- 539 8. LeCun, Y., Bengio, Y. & Hinton, G. Deep learning. *Nature* **521**, 436–444, [10.1038/nature14539](https://doi.org/10.1038/nature14539) (2015).
- 540 9. Goodfellow, I. J. *et al.* Generative adversarial networks. *arXiv preprint arXiv:1406.2661* [10.48550/arXiv.1406.2661](https://doi.org/10.48550/arXiv.1406.2661)  
541 (2014).
- 542 10. Isasa, I. *et al.* Comparative assessment of synthetic time series generation approaches in healthcare: leveraging patient  
543 metadata for accurate data synthesis. *BMC Med. Informatics Decis. Mak.* **24**, Article 27 (2024).
- 544 11. Morales, S. & Bowers, M. E. Time-frequency analysis methods and their application in developmental eeg data. *Dev.  
545 Cogn. Neurosci.* **54**, 101067, [10.1016/j.dcn.2022.101067](https://doi.org/10.1016/j.dcn.2022.101067) (2022).
- 546 12. Schumaker, L. L. *Spline Functions: Basic Theory* (Springer-Verlag, New York, 2007), 3rd edn.
- 547 13. Boor, C. D. *A Practical Guide to Splines* (Springer-Verlag, New York, 2001).
- 548 14. Brophy, E., Wang, Z., She, Q. & Ward, T. Generative adversarial networks in time series: A systematic literature review.  
549 *ACM Comput. Surv.* **55**, Article 199, [10.1145/3559540](https://doi.org/10.1145/3559540) (2023).
- 550 15. Ibarra-Fiallo, J. & Lara, J. A. Contextual deep learning approaches for time series reconstruction. In *2024 IEEE  
551 International Conference on Omni-Layer Intelligent Systems, COINS 2024* (Institute of Electrical and Electronics Engineers  
552 Inc., London, United Kingdom, 2024).
- 553 16. Yasuda, Y. & Onishi, R. Spatio-temporal super-resolution data assimilation (srda) utilizing deep neural networks with  
554 domain generalization. *J. Adv. Model. Earth Syst.* **15**, [10.1029/2023MS003658](https://doi.org/10.1029/2023MS003658) (2023).
- 555 17. Priessner, M. *et al.* Content-aware frame interpolation (cafi): deep learning-based temporal super-resolution for fast  
556 bioimaging. *Nat. Methods* **21**, 322–330, [10.1038/s41592-023-02138-w](https://doi.org/10.1038/s41592-023-02138-w) (2024).
- 557 18. Qiao, C. *et al.* A neural network for long-term super-resolution imaging of live cells with reliable confidence quantification.  
558 *Nat. Biotechnol.* [10.1038/s41587-025-02553-8](https://doi.org/10.1038/s41587-025-02553-8) (2025).
- 559 19. O’Shea, T. J. & West, N. Radio machine learning dataset generation with GNU radio. In *Proceedings of the GNU Radio  
560 Conference*, vol. 1 (2016).
- 561 20. DeepSig. Datasets (including radioml 2016.10a). <https://www.deepsig.ai/datasets/>. Accessed 2026-01-13.
- 562 21. DeepSig. Radioml 2018.01a dataset. <https://www.deepsig.ai/datasets/>. Accessed 2026-01-13.
- 563 22. McSharry, P. E., Clifford, G. D., Tarassenko, L. & Smith, L. A. A dynamical model for generating synthetic electrocardio-  
564 gram signals. *IEEE Transactions on Biomed. Eng.* **50**, 289–294, [10.1109/TBME.2003.808805](https://doi.org/10.1109/TBME.2003.808805) (2003).
- 565 23. McSharry, P. & Clifford, G. D. ECGSYN: A realistic ecg waveform generator (physionet). [https://physionet.org/physiotools/egsyn/](https://physionet.org/physiotools/<br/>566 egsyn/). Accessed 2026-01-13.
- 567 24. Krol, L. R., Pawlitzki, J., Lotte, F., Gramann, K. & Zander, T. O. Sereega: Simulating event-related eeg activity. *J.  
568 Neurosci. Methods* **309**, 13–24, [10.1016/j.jneumeth.2018.08.001](https://doi.org/10.1016/j.jneumeth.2018.08.001) (2018).
- 569 25. Pinceti, A., Sankar, L. & Kosut, O. Generation of synthetic multi-resolution time series load data. *arXiv:2107.03547*  
570 (2021).
- 571 26. Yuan, Z., Jiang, Y., An, Z., Ma, W. & Wang, Y. Seismic resolution improving by a sequential convolutional neural network.  
572 *PLOS ONE* **19**, e0304981, [10.1371/journal.pone.0304981](https://doi.org/10.1371/journal.pone.0304981) (2024).
- 573 27. Ibarra-Fiallo, J., Lara, J. A. & Agudelo Moreno, D. Cosibd, [10.5281/zenodo.15138853](https://doi.org/10.5281/zenodo.15138853) (2025). Version v1. Dataset.
- 574 28. Welch, P. D. The use of fast fourier transform for the estimation of power spectra: A method based on time averaging  
575 over short, modified periodograms. *IEEE Transactions on Audio Electroacoustics* **15**, 70–73, [10.1109/TAU.1967.1161901](https://doi.org/10.1109/TAU.1967.1161901)  
576 (1967).

- 577 29. Rabiner, L. R. & Gold, B. *Theory and Application of Digital Signal Processing* (Prentice Hall, 1975).
- 578 30. Marple, S. L., Jr. *Digital Spectral Analysis with Applications* (Prentice Hall, 1987).
- 579 31. Kuleshov, V., Enam, S. Z. & Ermon, S. Audio super resolution using neural networks. *arXiv preprint arXiv:1708.00853*  
580 [10.48550/arXiv.1708.00853](https://arxiv.org/abs/1708.00853) (2017).
- 581 32. Kaniraja, C. P., M, V. D. & Mishra, D. A deep learning framework for electrocardiogram (ecg) super resolution and  
582 arrhythmia classification. *Res. on Biomed. Eng.* **40**, 199–211, [10.1007/s42600-023-00320-x](https://doi.org/10.1007/s42600-023-00320-x) (2024).
- 583 33. Luciw, M. D., Jarocka, E. & Edin, B. B. Multi-channel eeg recordings during 3,936 grasp and lift trials with varying  
584 weight and friction. *Sci. Data* **1**, 140047, [10.1038/sdata.2014.47](https://doi.org/10.1038/sdata.2014.47) (2014).
- 585 34. Yamagishi, J., Veaux, C. & MacDonald, K. CSTR VCTK Corpus: English multi-speaker corpus for CSTR voice cloning  
586 toolkit (version 0.92), [10.7488/ds/2645](https://doi.org/10.7488/ds/2645) (2019).
- 587 35. Forestier, G., Petitjean, F., Dau, H. A., Webb, G. I. & Keogh, E. Generating synthetic time series to augment sparse datasets.  
588 In *2017 IEEE International Conference on Data Mining (ICDM)*, 865–870, [10.1109/ICDM.2017.106](https://doi.org/10.1109/ICDM.2017.106) (IEEE, 2017).

## 589 Acknowledgments

590 This research was supported by Dean's Office of the Polytechnic College of the San Francisco de Quito University and partially  
591 by ProyExcel-0069 project of the Andalusian University, Research and Innovation Department.

## 592 Author Contributions

593 J. I. F. handled the methodological design for artificial data creation, probabilistic analysis, spline-based variations, noise  
594 distributions, and random node selection. J. A. L. was responsible for the time series methodological design. D. A. M.  
595 performed data processing and validation analysis. All of the authors have contributed to writing the manuscript.

## 596 Competing Interests

597 The authors declare no competing interests