

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

Julio Ibarra-Fiallo¹, D'hamar Agudelo-Moreno¹, and Juan A. Lara²

⁴ ¹Colegio de Ciencias e Ingenierías, Universidad San Francisco de Quito, Cumbayá, Ecuador

⁵ ²Universidad de Córdoba, Córdoba, España

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

ABSTRACT

The increasing application of time-series analysis in fields like biomedical engineering or telecommunications emphasizes the need for high-quality data to train and evaluate advanced machine learning models. Acquiring real-world temporal data at suitable resolutions is often limited by ethical, economic, or practical constraints. We introduce CoSiBD (Complex Signal Benchmark Dataset for Super-Resolution), a synthetic dataset designed for reproducible time-series super-resolution research. CoSiBD provides 2,500 high-resolution signals ($N = 5,000$ samples each over a reference domain $\tau \in [0, 4\pi]$) with aligned low-resolution versions at four levels (150, 250, 500, and 1,000 samples) obtained via uniform decimation. Signals are generated with diverse non-stationary behaviors through piecewise frequency modulation and spline-based amplitude envelopes, and provides both clean and noisy variants. Signals are distributed as NumPy arrays, plain text, and JSON, with comprehensive metadata describing segment structure, generation parameters, and seeds for full reproducibility. Technical validation of the dataset analyzes spectral properties and reports baseline SR benchmarking and transfer experiments on EEG and speech to illustrate utility on real-world signals super-resolution.

Background & Summary

The analysis and simulation of temporal signals are fundamental across science and engineering. These techniques provide critical insights into dynamic processes in multiple domains. For instance, in biomedical research, physiological time series such as electroencephalography (EEG) and electrocardiography (ECG) support the study of neural and cardiac function¹⁻³. Another example is telecommunications, which rely on signal processing to ensure data fidelity across noisy media⁴. Developing robust tools for interpreting time-varying data continues to support both scientific discovery and practical applications.

Recent advances in deep learning have contributed significantly to this field by enabling automatic extraction of complex features from raw signals. Deep learning approaches, such as Convolutional Neural Networks (CNNs) or Generative Adversarial Networks (GANs) have demonstrated improved performance over traditional techniques in image, speech, and time-series processing tasks^{5,6}. These models support fine-grained signal reconstruction, allowing researchers to explore temporal dynamics in new ways.

Despite this progress, deep learning methods for temporal signal processing often require large quantities of labeled, high-quality data. Access to such data is frequently constrained. For instance, in medicine, data access is constrained by medical privacy regulations such as the General Data Protection Regulation (GDPR) and the Health Insurance Portability and Accountability Act (HIPAA)⁷. In other domains, including telecommunications, data availability is limited by proprietary protocols and the high cost of acquiring large-scale channel sounding datasets for diverse environments⁸. These limitations are particularly relevant in super-resolution (SR) tasks, where models require paired low- and high-resolution signals for effective training. Here, temporal SR refers to reconstructing a higher-sample-count (higher temporal resolution) discrete sequence from a uniformly decimated low-resolution version.

This task has broad potential. In biomedical monitoring and sensing, for instance, SR can help reconstruct higher-resolution physiological time series (e.g., EEG), potentially improving the analysis of subtle physiological irregularities³. SR also applies to audio/speech enhancement and telecommunications (e.g., neural audio upsampling and bandwidth extension, relevant to telephony and compression)^{9,10}.

Deep learning offers adaptive alternatives to these traditional methods. For instance, CNNs are capable of modeling spatio-temporal structure present in real datasets across these domains. Preliminary work on synthetic time-series generation indicates

34 potential for SR^{10,11}, but the lack of accessible, high-quality paired datasets remains a significant barrier to progress.

35 Synthetic datasets offer one solution to this problem, allowing researchers to design reproducible training environments
36 that reflect the characteristics of real-world signals. Prior studies have used synthetic data in domains such as biomedical
37 signal analysis¹² and wireless communications¹³, demonstrating that synthetic approaches can help simulate complexity while
38 avoiding legal and practical restrictions associated with real-world data.

39 To support research in super-resolution for time-series data, we present the Complex Signal Benchmark Dataset (CoSiBD).
40 CoSiBD is a synthetic dataset composed of time-series signals with variable resolution, frequency characteristics, and noise
41 levels. As it is designed to resemble real-world signals, our dataset is intended with a double purpose: a) to provide a resource
42 for training and evaluating deep learning SR models under controlled, reproducible conditions, which can constitute a sort
43 of benchmark for this problem; and b) a resource for training deep learning models to be used for SR of real-world signals
44 (either directly or finetuning them with the real data), particularly in scenarios where real signals are scarce and not enough
45 for a complete training of those models. It includes non-stationary, piecewise-structured signals (via non-uniform interval
46 partitioning with change-points), multiple levels of resolution and noise, a technical validation suite, and publicly available
47 Python code to facilitate use. CoSiBD has been previously used in research presented at the International Conference on Signal
48 Processing and Machine Learning¹⁰, with good preliminary results for signal reconstruction using deep learning. CoSiBD is
49 made available to support further development in deep learning approaches for temporal super-resolution.

50 To further position CoSiBD with respect to existing public synthetic time-series resources, we summarize in the next subsection
51 representative datasets and simulators and highlight the practical gap addressed by our approach.

52 Related synthetic time-series resources

53 Publicly available synthetic resources for temporal signals exist, but they are typically designed for tasks other than time-series
54 SR, or they target a specific domain. In wireless communications, the RadioML family provides large collections of synthetic
55 complex I/Q sequences with varying SNR (Signal-to-noise ratio) and channel impairments, mainly to benchmark automatic
56 modulation classification rather than paired SR reconstruction^{13–15}. In biomedical signal processing, physiological simulators
57 such as ECGSYN (electrocardiography) and SEREEGA (EEG) enable controlled generation with tunable morphology, sampling
58 settings, and noise, supporting method development when real data access is constrained^{12,16,17}. In power systems, LoadGAN
59 provides multi-resolution generation of load time series across sampling rates and time horizons (from sub-second to long-term
60 scales), but it is not distributed as a standardized paired SR benchmark¹⁸. Domain-specific paired low-/high-resolution training
61 data can also be produced via physical forward modeling, e.g., low- and high-resolution 1D seismic traces for learning-based
62 resolution enhancement¹⁹.

63 Table 1 provides a concise view of representative public synthetic resources for temporal signals and clarifies how they relate to
64 the specific needs of time-series super-resolution (SR). We summarize each resource by its target *Domain* and *Form* (fixed
65 dataset vs. simulator/generator), and we highlight practical properties that determine whether a resource can be used as an
66 SR benchmark: whether it provides *paired LR–HR data*, whether it supports *multi-resolution* generation, what *noise/artifact*
67 mechanisms are available, and the *reproducibility granularity* (dataset-level variability vs. per-signal deterministic regeneration).

68 Overall, existing resources offer important strengths but also exhibit limitations when viewed through an SR-benchmark lens.
69 Domain-specific datasets such as RadioML provide large-scale controlled variability (e.g., SNR and channel impairments) and
70 are highly valuable for tasks like modulation classification, but they are not distributed as paired LR–HR SR targets. On the
71 other hand, physiological simulators (e.g., ECGSYN and SEREEGA) enable flexible generation with tunable morphology,
72 sampling settings, and noise, which in principle can be used to create LR–HR pairs; however, they typically do not provide
73 a standardized, fixed paired benchmark designed explicitly for SR evaluation across multiple difficulty levels. Similarly,
74 multi-resolution generators in other domains may capture realistic structure at multiple sampling rates, but often lack an aligned,
75 reproducible LR–HR pairing protocol and a unified dataset release that facilitates direct, comparable benchmarking across
76 methods.

77 These observations motivate the need for a public resource that combines: (i) fixed, aligned LR–HR pairs for SR, (ii) controllable
78 nuisance modeling (noise and structured interference), and (iii) strong reproducibility at the per-signal level, while remaining
79 broad enough to support benchmarking and transfer across domains.

80 CoSiBD is designed to close the gap by providing a fixed, standardized set of LR–HR pairs with explicit nuisance modeling
81 (noise and structured interference) and comprehensive metadata, enabling reproducible benchmarking across multiple difficulty
82 levels (defined by varying downsampling factors and noise intensities).

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 ^{13,14}	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 ¹⁵	Wireless communications (I/Q)	Dataset	No (classification benchmark)	N/A (not SR)	Simulated channel effects + SNR variability	Dataset-level; not SR-paired
ECGSYN ^{12,16}	ECG (physiology)	Simulator/tool	Configurable	Configurable (via sampling settings)	Model-based; supports controlled variability	Configurable via simulator parameters (user-defined)
SEREEGA ¹⁷	EEG (physiology)	Simulator/toolbox	Configurable	Configurable (user-defined)	Supports noise and event-related components	Configurable via simulator parameters (user-defined)
LoadGAN ¹⁸	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) ¹⁹	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.

83 Methods

84 The methodology used to generate the synthetic temporal signals that constitute the CoSiBD dataset is illustrated in Figure 1,
 85 and will be explained later. The signal generation process is designed to produce time series exhibiting structural properties
 86 commonly observed in real-world temporal data, including variable frequency content, smooth transitions, and intermittent
 87 high-frequency activity. A key aspect of the procedure is the generation of signals at multiple temporal resolutions, enabling the
 88 construction of paired datasets for super-resolution (SR) benchmarking.

89 **Design rationale inspired by real signals.** It is important to note that one of the main applications of the proposed dataset will
 90 be the training of deep learning models to be used for SR purposes in other real-world datasets, like, for instance, physiological
 91 or speech signals. Therefore, there is a need for our dataset to resemble real-world data. In particular, real signals exhibit (i)
 92 non-stationary regime changes, (ii) coexisting low- and high-frequency components with intermittent transients, (iii) smooth
 93 amplitude-envelope evolution, and (iv) slow baseline drift and measurement noise. CoSiBD instantiates these properties
 94 via non-uniform interval partitioning with change-points, separate low/high-frequency bands, spline-based envelopes and
 95 frequency profiles, and explicit offset/noise terms. Figure 2 provides qualitative examples of these motivating properties found
 96 in real-world time series; the main goal of our dataset is to be able to capture challenging structure for SR benchmarking rather
 97 than match a specific domain distribution.

98 The design of CoSiBD is informed by the inherent limitations of standard synthetic datasets in capturing the irregular
 99 morphologies of real-world signals. As shown in Figure 2, physiological signals such as ECG or EEG (**a–b**) are rarely
 100 stationary, often exhibiting abrupt regime shifts and highly structured spectral components that pose significant challenges for
 101 temporal reconstruction. By implementing explicit change-points and multi-band frequency partitioning, CoSiBD attempts to
 102 approximate these localized dynamics, providing a more demanding benchmark than globally stationary models.

103 Similarly, the inclusion of spline-based modulation seeks to address the continuous but non-linear fluctuations observed in
 104 acoustic data (**c–d**). The amplitude envelopes and F0 trajectories in speech represent a level of temporal complexity that
 105 rigid parametric approaches often oversimplify. Using smooth spline profiles allows the framework to emulate these high-
 106 order dependencies without assuming a fixed functional form. This ensures that the resulting dataset, while still synthetic,
 107 exposes super-resolution models to the morphological variability—such as pitch drifting and dynamic scaling—required for
 108 generalization to authentic signals.

109 The signal generation pipeline involves the following steps:

- 110 **1. Base frequency band definition:** A set of distinct frequency bands is defined to represent the underlying spectral content
 111 of the signals. These can be adjusted to reflect application-specific characteristics.
- 112 **2. Non-uniform interval partitioning:** The total signal duration is divided into multiple intervals of variable length. The
 113 interval lengths are determined probabilistically to introduce variability in the signal structure.

CoSiBD Dataset Generation Process

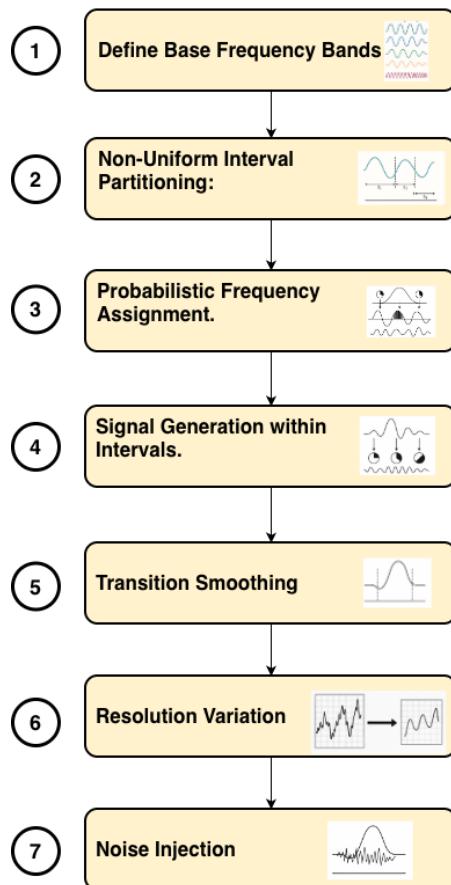
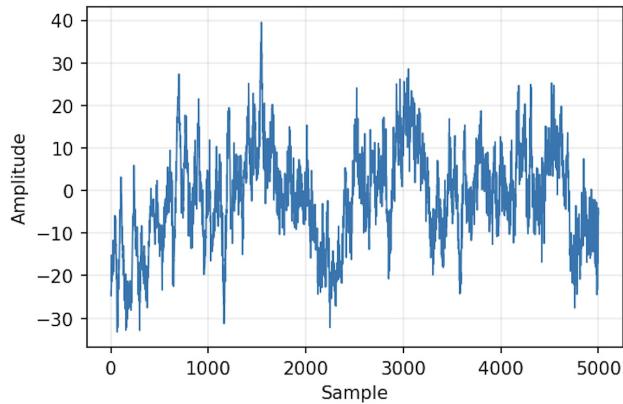
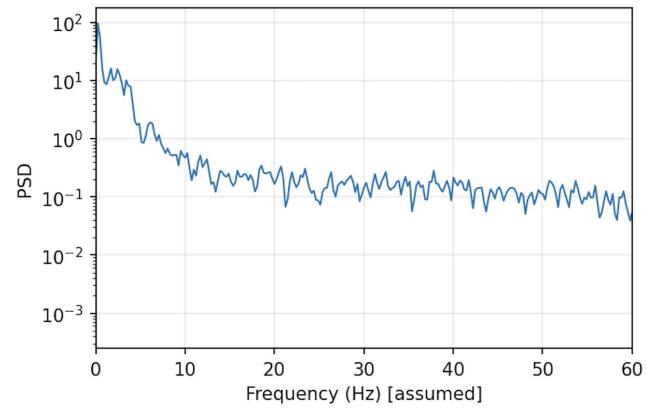


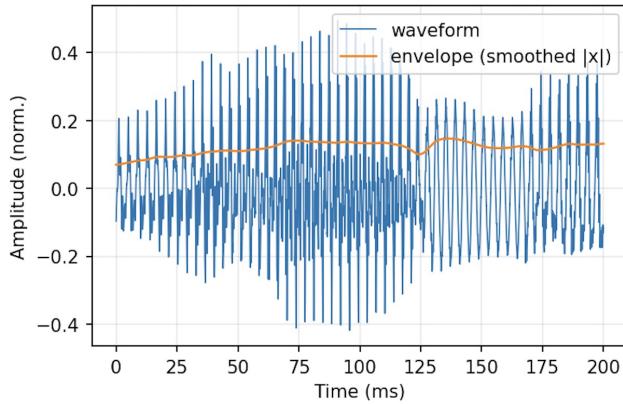
Figure 1. Schematic overview of the CoSiBD signal generation process.



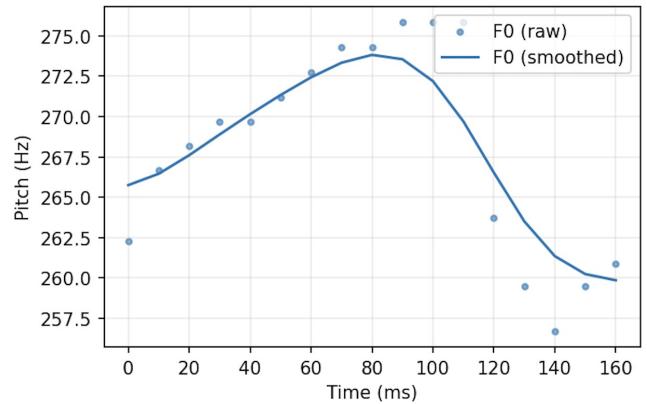
(a) Physiological example: ECG/EEG like waveform



(b) Physiological PSD (Welch, qualitative).



(c) Speech example: waveform + amplitude envelope



(d) Speech example: pitch trend (F0)

Figure 2. Qualitative real-signal properties motivating the CoSiBD design, specifically non-stationarity and spectral structure in physiological data, and envelope-pitch dynamics in speech.

- 114 3. **Frequency assignment:** Each interval is assigned a dominant frequency band, sampled according to a predefined
 115 probability distribution. This introduces spectral variation over time.
- 116 4. **Signal synthesis:** A sinusoidal waveform, or a combination of sinusoids within the assigned frequency band, is generated
 117 for each interval. Signal parameters such as amplitude and phase are configurable.
- 118 5. **Transition smoothing:** To avoid discontinuities at interval boundaries, a smoothing function is applied to overlapping
 119 segments. This ensures gradual transitions between intervals with different frequency content.
- 120 6. **Resolution variation:** All signals are initially synthesized at a high temporal resolution (5,000 samples over the domain
 121 $[0, 4\pi]$). Lower-resolution versions are created using simple decimation (uniform subsampling). This keeps the SR task
 122 aligned with reconstructing the original high-resolution target; the low-resolution observation is obtained by subsampling
 123 the original sequence without pre-filtering. Reconstructing low-pass filtered signals is not an objective of CoSiBD.
 124 For reproducibility, given a high-resolution sequence $x_{\text{HR}}[n]$ of length $N = 5000$ and a target low-resolution length
 125 $M \in \{1000, 500, 250, 150\}$, we form $x_{\text{LR}}[i] = x_{\text{HR}}[n_i]$ using the fixed index set $n_i = \left\lfloor \frac{i(N-1)}{M-1} + 0.5 \right\rfloor$ for $i = 0, \dots, M-1$
 126 (applied identically to the time array). This reduces to standard stride decimation when M divides N .
- 127 7. **Noise injection:** Controlled levels of synthetic noise are added to the signals to emulate different data acquisition
 128 scenarios. Two noise types are implemented: (1) Additive white Gaussian noise (AWGN) with configurable standard
 129 deviation (relative to signal RMS amplitude), representing broadband sensor thermal noise; and (2) structured sinusoidal
 130 noise bursts (deterministic sinusoidal components), representing narrow-band interference such as powerline hum
 131 (50/60 Hz). Noise is applied probabilistically (in the released dataset, `noise_profile` records `p_has_noise=0.5`
 132 per signal, and the realized subset with noise is approximately half). When noise is present, the specific parameters
 133 (type, amplitude, frequency for structured noise) are recorded in the metadata, allowing users to benchmark denoising or
 134 super-resolution under specific degradation conditions.

135 **Rationale for structured 50/60 Hz interference and noise.** Real measurement pipelines frequently contain narrow-band
 136 interference (e.g., mains hum) superimposed on broadband sensor noise. To reflect this common acquisition artifact, CoSiBD
 137 includes an optional structured sinusoidal component in addition to Gaussian noise. CoSiBD signals are generated over a
 138 reference domain (by default $\tau \in [0, 4\pi]$); interpreting τ as physical time (and therefore reporting frequencies in Hz) requires
 139 an explicit time scaling. Throughout this manuscript we adopt an illustrative convention that maps the reference domain to a
 140 duration $T = 4\pi$ seconds, under which the structured component can be interpreted as a 50/60 Hz-like powerline interference
 141 term, while the broadband term represents the measurement noise floor.

142 Figure 3 illustrates this qualitative motivation. Subfigure (a) displays a clean synthetic ECG signal, serving as a baseline.
 143 Subfigure (b) shows the same signal contaminated with a 50 Hz sinusoidal component (mimicking powerline interference
 144 in Europe/Asia), while (c) illustrates the effect of a 60 Hz component (typical of the Americas). Subfigure (d) presents the
 145 frequency spectrum comparison, clearly showing the narrow-band interference spikes associated with these powerline artifacts.
 146 This explicitly models the periodic contamination observed in real recordings²⁰; the intent is not to reproduce a specific device
 147 transfer function but to include realistic nuisance factors that SR models must handle.

148 **Sampling units and frequency interpretation.** CoSiBD signals are provided as discrete sequences $x[n]$ (e.g., $N = 5,000$
 149 samples) that are directly used as inputs/targets by SR models. The internal generation domain $\tau \in [0, 4\pi]$ is a reference
 150 parameterization; interpreting it as physical time requires choosing a duration T (in seconds) for the reference interval.
 151 Under this convention, the implied sampling rate is $f_s = N/T$ and all frequencies reported in Hz scale linearly with $4\pi/T$.
 152 Throughout this manuscript, when reporting example frequencies in Hz we adopt the illustrative convention $T = 4\pi$ s, yielding
 153 $f_s \approx 5000/(4\pi) \approx 398$ Hz; other equally valid mappings exist depending on application. Consequently, any band-specific
 154 interpretation in Hz (e.g., “low/high” frequency ranges) should be understood under the chosen T ; changing T rescales all
 155 reported Hz values while preserving the underlying discrete sequences, which is a key feature of CoSiBD’s design.

156 Figure 4 clarifies this convention using three panels (a–c). Panel (a) shows the normalized frequency spectrum (cycles/sample),
 157 which is intrinsic to the discrete sequence. Panels (b) and (c) illustrate how this same spectrum maps to physical units (Hz) under
 158 different assumed sampling rates: (b) assumes $f_s \approx 397.9$ Hz (corresponding to $T = 4\pi$ s), while (c) assumes $f_s \approx 1000$ Hz.
 159 This demonstrates that while the discrete data remains identical, the physical interpretation scales with the user-defined duration.

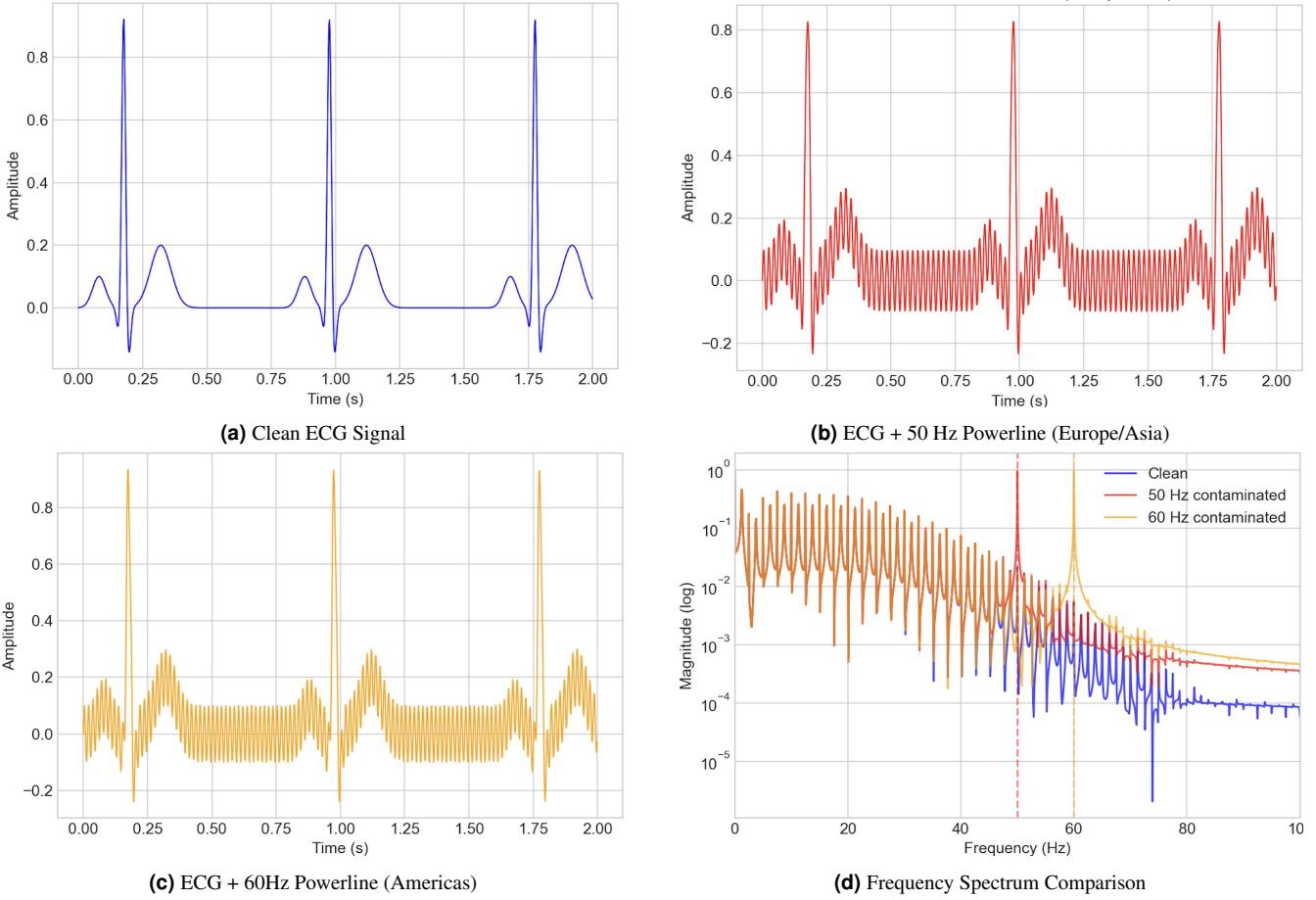


Figure 3. Qualitative motivation for the structured interference term used in CoSiBD.

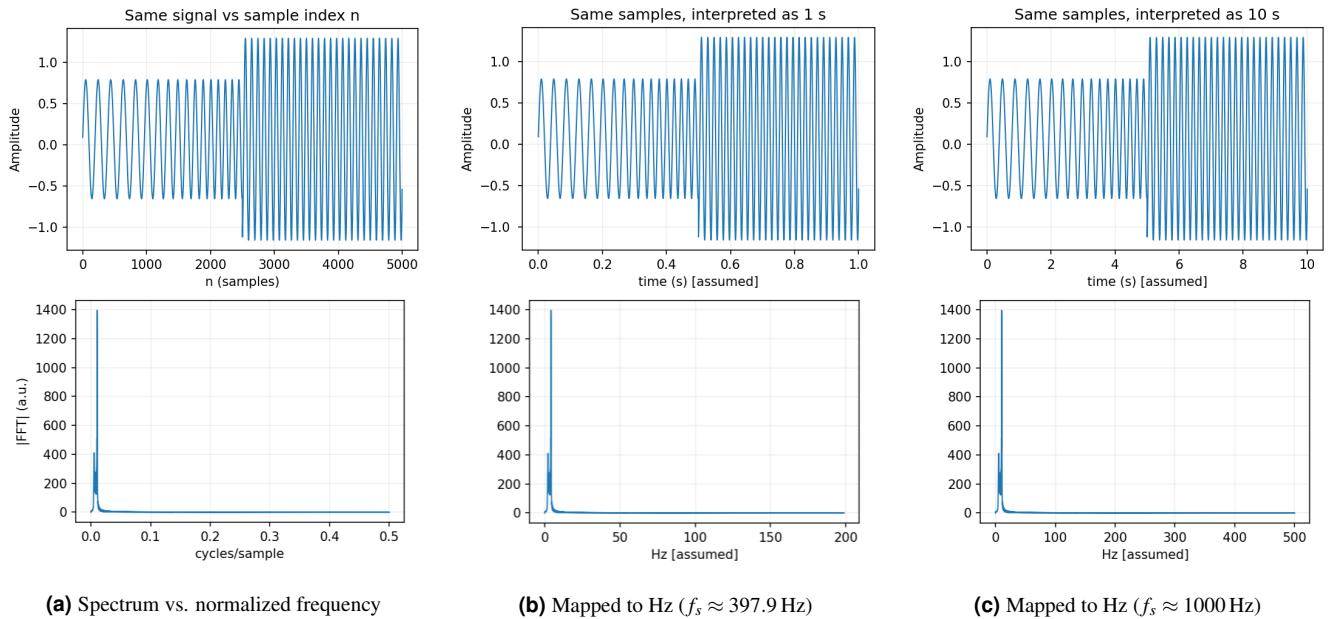


Figure 4. Frequency domain mapping under different sampling rate assumptions.

160 Data Records

161 The Complex Signal Benchmark Dataset (CoSiBD) is publicly available on Zenodo²¹ and consists of synthetic temporal signals,
162 mainly created to support the development and evaluation of temporal super-resolution (SR) algorithms, and also to train deep
163 learning models that can be used for SR in real-world signals. This section provides an overview of the dataset structure,
164 content, and storage format, as well as the parameters that rule the generation of data and the metadata that enrich our dataset.

165 The dataset comprises 2,500 high-resolution signals, each with corresponding subsampled versions at four resolution levels,
166 organized into three main categories:

- 167 • **High-resolution signals:** 2,500 signals with 5,000 samples each, spanning the domain $T = [0, 4\pi]$ (s), which, under the
168 illustrative convention used, corresponds to $f_s = 5000/(4\pi) \approx 398$ Hz. Each signal is stored in three formats: NumPy
169 compressed format (.npz), plain text (.txt), and JSON (.json). Per-signal metadata (frequency profiles with explicit
170 change-points (`base_points` and `high_freq_points`) and segment labels (`variation_type`), amplitude
171 envelopes, spline parameters, vertical offsets, noise configurations, and seeds) is provided in a consolidated JSON file
172 (`signals_metadata.json`) with one entry per signal, enabling exact regeneration.
- 173 • **Simple subsampled signals:** Uniform decimation (uniform subsampling) of each signal to four target resolutions: 150
174 (illustrative $f_s \approx 11.9$ Hz for $T = 4\pi$ s), 250 (illustrative $f_s \approx 19.9$ Hz for $T = 4\pi$ s), 500 (illustrative $f_s \approx 39.8$ Hz for
175 $T = 4\pi$ s), and 1,000 samples (illustrative $f_s \approx 79.6$ Hz for $T = 4\pi$ s). These low-resolution versions serve as inputs for
176 SR benchmarking against the original 5,000-sample target. Stored in .npz, .txt, and .json formats.

177 The dataset is provided as consolidated files under `SignalBuilderC/data/`. High-resolution signals are stored as
178 `signals_high_resolution_5000.[npz|txt|json]`. Simple subsampled (decimated) signals are stored as `signals_`
179 `subsampled_simple_{150,250,500,1000}.[npz|txt|json]`. Dataset-level metadata (described later in de-
180 tail) and configuration are stored in `signals_metadata.json` (per-signal metadata, one entry per signal), `signals_`
181 `metadata Consolidated_2500.json`, and `dataset_summary.json`.

182 Regarding the three formats used for both high-resolution and subsampled signals, we provide here some additional information
183 for each format: (1) NumPy compressed format (.npz) containing the signal array, time array, and (for high-resolution only)
184 clean signal without noise; (2) consolidated plain text format (.txt) with one signal per row (samples separated by whitespace)
185 for maximum portability; and (3) JSON format (.json) with both time and signal arrays for web-based applications and
186 interoperability.

187 Reproducibility is ensured through documented seeds: each high-resolution signal is generated using a unique seed (ranging
188 from 10,000 to 12,499), enabling exact regeneration of individual signals or the entire dataset. All generation parameters
189 (described later in detail) are stored in metadata JSON files, including: (1) frequency profile parameters—`tau_frequency`
190 values from uniform distribution [1, 2] with 0.05 step; (2) amplitude envelope parameters—`amplitude_spline_type` is
191 predominantly `zero_order` (approximately 70%), otherwise `tension` with `tau_amplitude` drawn from {1, 3, 5, 8, 10, 12, 15,
192 20}; (3) vertical offsets—normally distributed (`mean=0`, `SD=3.0`); and (4) noise configurations—per-signal `noise_profile`
193 with `p_has_noise=0.5` (and `p_gaussian=0.5` when noise is present).

194 Metadata schema and example

195 CoSiBD provides per-signal metadata to support (i) deterministic regeneration, (ii) principled partitioning (e.g., by noise
196 type/level or segment labels), and (iii) analysis of the piecewise structure induced by change-points. Table 2 summarizes
197 representative fields contained in `signals_metadata.json`. A minimal example entry is shown below (one signal; values
198 truncated for brevity).

199 Table 2 details the key fields describing each signal, specifically the `variation_type` list which provides categorical
200 labels aligned with the `base_points` anchors (e.g., “low”, “high”, or “no_change”), enabling users to filter for signals
201 containing specific dynamic behaviors. The following JSON snippet shows a simplified illustrative metadata record (Example
202 1; values truncated for readability). Figure 5 provides a visual counterpart of the same record: it highlights how consec-
203 utive `base_points` (and, when present, `high_freq_points`) define temporal intervals and frequency targets, while
204 `variation_type` provides the corresponding per-anchor labels (here, yielding two intervals: “low” then “high”), and
205 `amp_knots/amp_values` define the amplitude envelope.

```
206 {  
207     "t_start": 0.0,  
208     "t_end": 12.566370614359172,
```

```

209 "fs_high": 397.88735772973837,
210 "tau_frequency": 1.15,
211 "amplitude_spline_type": "zero_order",
212 "tau_amplitude": "N/A",
213 "vertical_offset": 0.06905161748158965,
214 "base_points": [[0.0, 1.0], [6.28, 3.0], [12.56, 3.0]],
215 "high_freq_points": [[0.0, 0.0], [6.28, 0.0], [12.56, 0.0]],
216 "variation_type": ["low", "high", "high"],
217 "amp_knots": [0.0, 6.28, 12.56],
218 "amp_values": [0.72, 1.22, 0.96],
219 "noise_profile": {"has_noise": true, "noise_type": "gaussian", "p_has_noise": 0.5, ...},
220 "seed": 10000,
221 "signal_id": "signal_0000",
222 "index": 0
223 }

```

Field	Type	Example	Meaning
signal_id	String	"signal_0000"	Unique identifier for the generated signal.
index	Integer	0	Numeric index of the signal (0 to N-1).
seed	Integer	10000	Seed used for deterministic generation of this specific signal.
t_start, t_end	Float	0.0, 12.56	Start and end time of the signal domain.
base_points	Array	[[0.0, 2.07]...]	Control points (t, f) for the frequency spline.
high_freq_points	Array	[[0.0, 0.0]...]	Control points (t, f) for an optional high-frequency component.
variation_type	List[Str]	["low", "high", "high"]	Categorical labels aligned with base_points anchors (same length), describing the intended band at each change-point (e.g., "low", "high", "no_change"); interval labels are derived between successive anchors.
tau_frequency	Float	1.15	Tension parameter for the frequency spline.
amplitude_spline_type	String	"zero_order"	Interpolation family for the amplitude envelope (e.g., step-wise or tension spline).
tau_amplitude	Float / String	10.0	Tension parameter for the amplitude spline (or "N/A" for zero-order).
amp_knots	Array	[0.0, 6.28, 12.56]	Knot times defining the amplitude envelope segments.
amp_values	Array	[0.72, 1.22, 0.96]	Amplitude values at knots (piecewise-constant or spline-interpolated).
noise_profile	Dict	{...}	Parameters for added noise (if any).

Table 2. Metadata schema describing the JSON structure for each signal. Key fields define the frequency trajectory (base_points, variation_type) and amplitude envelope, allowing precise reconstruction or filtering.

224 Parameters for signal generation

225 As we anticipated before, our dataset is generated from a set of configuration parameters summarized in Table 3. Each
226 high-resolution signal was generated with a unique seed (10,000–12,499) and sampled parameter values within the defined
227 ranges, supporting diversity while maintaining reproducibility. Parameters marked as *Sampled (seed-controlled)* are drawn
228 per-signal using the recorded seed, while fixed defaults are held constant across the dataset. Unless otherwise stated, these
229 defaults correspond to the settings used in the Zenodo release, and the CLI follows the same behavior unless a user explicitly
230 overrides a parameter.

231 Concretely, generation proceeds as a deterministic pipeline conditioned on the seed: (i) the seed initializes all pseudo-random
232 draws; (ii) a set of frequency-trajectory and envelope parameters is sampled within the reported ranges; (iii) a continuous-time
233 signal is synthesized over the fixed domain ($T = 4\pi$ s convention) by combining a time-varying frequency trajectory with
234 an amplitude envelope; and (iv) optional nuisance terms (offset and additive noise) are applied. Therefore, fixing the seed
235 fully determines the sampled parameters and the resulting signal instance, which supports exact reproduction and controlled
236 ablations.

237 **Frequency trajectory.** The overall (instantaneous) frequency evolution is controlled by the sampled low/high frequency bands
238 ("Low Frequency" and "High Frequency"), the number of transitions ("Change Points"), and their temporal placement ("Change
239 Locations"). The categorical "Variation Type" ({low, high, no_change}) provides labels aligned with the base_points
240 anchors, indicating whether the trajectory targets the low band, the high band, or remains stable around each change-point;

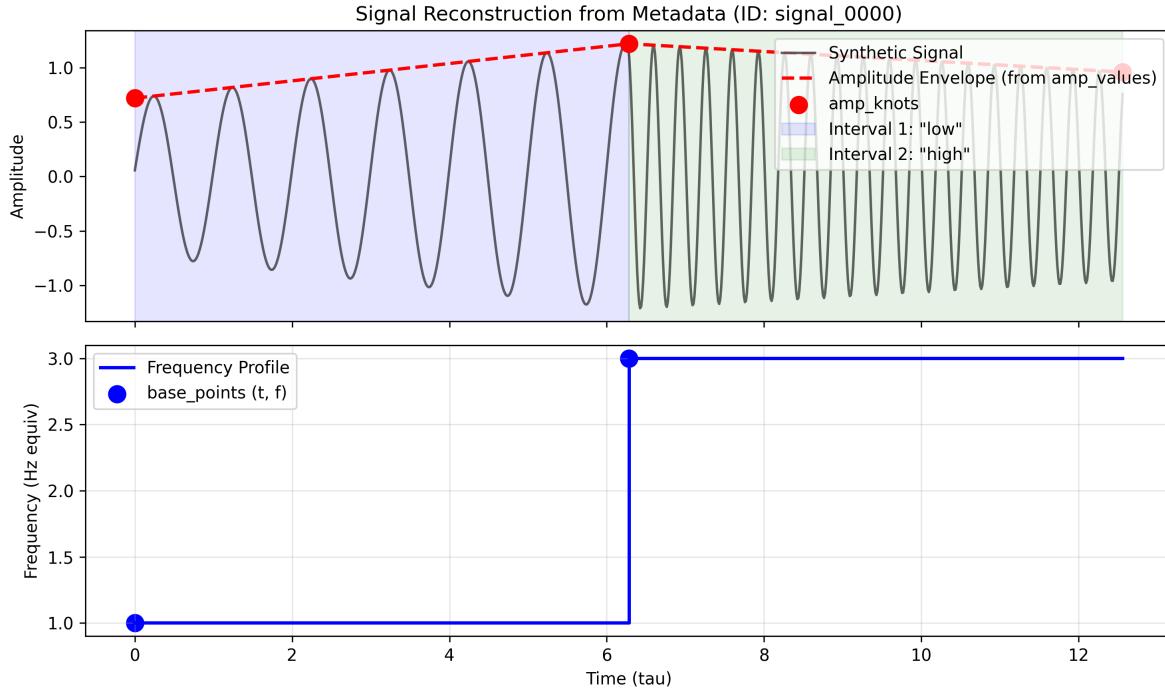


Figure 5. Visual representation of an illustrative metadata record for signal_0000 (Example 1). The plot illustrates how `base_points` define temporal intervals and frequency targets, while `amp_values` control the amplitude envelope segments, corresponding directly to the JSON structure.

241 this supports controlled balancing and downstream filtering via metadata. The smoothness/shape of the resulting trajectory is
 242 governed by the spline settings, including `tau_frequency` (“Tension (freq)”).

243 **Amplitude envelope.** Independently of the frequency trajectory, the signal amplitude is modulated by an envelope whose
 244 overall scale is constrained by “Amplitude Range”. The envelope shape is generated via the specified interpolation scheme
 245 (“Spline Type”) and, when applicable, by the sampled “Tension (amp)” parameter. This yields signals with comparable
 246 frequency content but different amplitude dynamics.

247 **Offsets and noise.** To emulate common acquisition artifacts without tying the generator to a specific instrument, we add a
 248 per-signal “Vertical Offset” and inject additive noise stochastically according to “Noise Prob.”. When noise is enabled for a
 249 given signal, its parameters are recorded in metadata to preserve reproducibility.

250 **Resolution and sampling.** Each instance is first generated at the target high-resolution length (e.g., 5,000 samples in the
 251 released dataset), and lower-resolution views are derived by subsampling, enabling consistent multi-scale analysis and SR
 252 benchmarking (Figure 6).

253 **Mathematical formulation.** Let $t \in [t_{\text{start}}, t_{\text{end}}]$ with the paper’s convention $T = t_{\text{end}} - t_{\text{start}} = 4\pi$ s. A generated signal can be
 254 written as

$$x(t) = v + a(t) \sin(\phi(t)) + \varepsilon(t), \quad (1)$$

255 where v is a vertical offset, $a(t) \geq 0$ is the amplitude envelope, and $\varepsilon(t)$ is an optional additive noise term. The instantaneous
 256 frequency trajectory is encoded through the phase derivative

$$f(t) = \frac{1}{2\pi} \frac{d\phi(t)}{dt}, \quad (2)$$

257 and is constructed from a spline defined by a set of control points and a tension parameter (Table 3). This separation (frequency
 258 trajectory vs. amplitude envelope) allows generating signals that share the same transition structure but differ in amplitude
 259 dynamics, or vice versa.

260 **Mapping Table 3 to metadata fields.** The parameters in Table 3 correspond to concrete entries in each per-signal metadata
261 record (Figure 5 shows an illustrative example):

- *Seed* → seed (and derived identifiers signal_id, index). Fixing seed reproduces all sampled values.
- *Domain $T = 4\pi$* convention → t_start, t_end.
- *Change points / locations / variation type* → base_points (frequency control points), variation_type (interval labels), and auxiliary arrays such as high_freq_points used to represent high-frequency targets during transitions.
- *Frequency tension* → tau_frequency.
- *Amplitude range / spline settings* → amplitude_spline_type, tau_amplitude, and the envelope knot representation amp_knots and amp_values.
- *Vertical offset* → vertical_offset.
- *Noise probability and settings* → noise_profile (including whether noise was applied and the sampled noise parameters for that signal).

272 Some fields (e.g., fs_high) are reported as part of the complete record to make downstream processing explicit; they are
273 derived from the chosen resolution and the domain convention.

Parameter	Range / Options	Default	Description
Low Frequency	1–5 Hz	Sampled (seed-controlled)	Low-frequency component ($T = 4\pi$ s convention).
High Frequency	20–100 Hz	Sampled (seed-controlled)	High-frequency variations for transitions.
Change Points	2–11	Sampled (seed-controlled)	Number of frequency transitions per signal.
Change Locations	Continuous	Sampled (seed-controlled)	Time locations where transitions occur.
Variation Type	{low, high, no_change}	Balanced	Category of frequency change per segment.
Amplitude Range	3–16	Sampled (seed-controlled)	Bounds for amplitude envelope generation.
Vertical Offset	$N(0, 3.0)$	0.0	Normally distributed offset added to signals.
Spline Type	Zero-Order (70%), Tension (30%)	-	Interpolation method for envelopes.
Tension (freq)	[1, 2]	1.5	Tension parameter for frequency splines.
Tension (amp)	{1, 3, 5, 8, 10, 12, 15, 20}	Sampled (seed-controlled)	Tension parameter for amplitude splines ("N/A" for zero-order).
Noise Prob.	0.0–1.0	0.5	Probability of adding noise to a signal.
Seed	Integer	-	Unique initialization seed per signal.

Table 3. Configuration parameters used to generate the dataset. Ranges define the sampling space for the 2,500 signals, ensuring diversity. When not explicitly set by the user, parameters are either sampled per signal (seed-controlled) or take fixed defaults as listed (defaults used for the Zenodo release).

274 To explicitly characterize dataset diversity and complexity, CoSiBD spans multiple controlled axes of variation (Table 3),
275 including the number and location of change points, categorical transition types, low/high frequency bands, and amplitude-
276 envelope configurations. While the dataset is synthetic and not fitted to match a single domain-specific distribution, these
277 controlled variations provide reproducible coverage of common real-world time-series phenomena such as non-stationarity,
278 transient high-frequency events, and additive noise.

279 Figure 6 shows a representative signal from the dataset sampled at different resolution levels, as well as a version with added
280 noise. This illustrates the variety of sampling and noise conditions included in CoSiBD. Figure 7 displays four additional
281 synthetic signals generated using different configuration parameters. These examples demonstrate the variability in temporal
282 structure across instances in the dataset.

283 Custom Dataset Generation

284 In addition to the pre-generated dataset, the CoSiBD package includes a command-line interface (CLI) that allows users to
285 generate custom datasets with their own parameter distributions. Figure 8 demonstrates the usage of this tool. The command
286 specifies the number of signals (-n_signals), the target high-resolution length in samples (-resolution, e.g., 5000),
287 and the probability of applying noise to each signal (-noise_prob, e.g., 0.5, meaning noise is injected into roughly half of
288 the signals). The output log confirms the resolved configuration (signal count, domain, and output directory), reports progress
289 during generation, and summarizes the created artifacts (per-signal files and the consolidated signals_metadata.json).

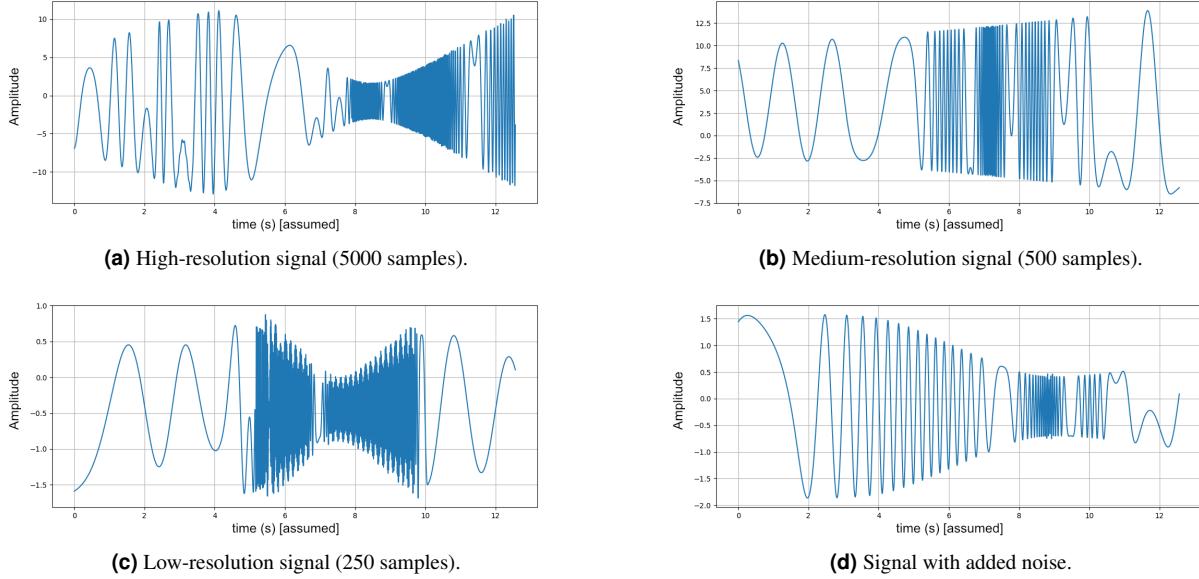


Figure 6. A synthetic signal sampled at different resolutions: (a) high (5000 samples), (b) medium (500 samples), (c) low (250 samples), and (d) with added noise. These examples reflect the multi-resolution and noise conditions present in the dataset.

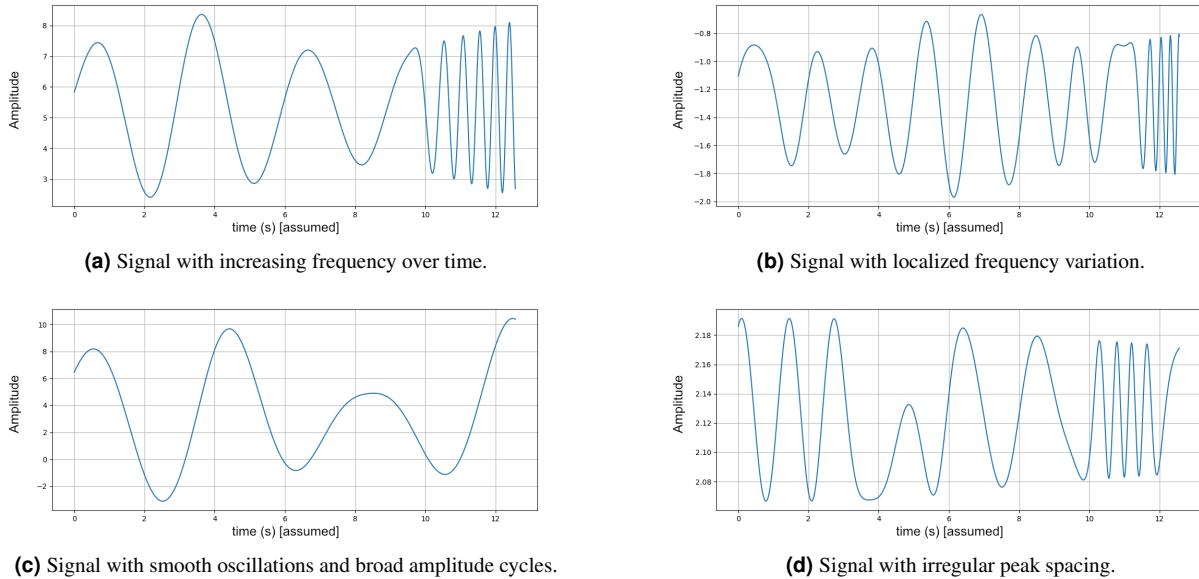


Figure 7. Examples of synthetic signals in the dataset generated with different parameter configurations. Each signal presents a distinct temporal profile.

```

$ python generate_dataset.py --n_signals 2500 --resolution 5000 --noise_prob 0.5

[INFO] Initializing Signal Generator...
[INFO] Configuration:
  - Signals: 2500
  - Resolution: 5000 samples
  - Domain: [0, 4pi]
  - Noise Probability: 0.5
[INFO] Output Directory: ./dataset_output/

Processing: 100% |██████████| 2500/2500 [00:45<00:00, 55.20it/s]

[SUCCESS] Dataset generation complete.
[INFO] Generated 2500 .npz files.
[INFO] Metadata saved to ./dataset_output/signals_metadata.json

$ ls -lh ./dataset_output/ | head -n 5
total 420M
-rw-r--r-- 1 user staff 120K Jan 10 10:00 signal_0000.npz
-rw-r--r-- 1 user staff 120K Jan 10 10:00 signal_0001.npz
-rw-r--r-- 1 user staff 120K Jan 10 10:00 signal_0002.npz

```

Figure 8. Demonstration of the CoSiBD Command Line Interface (CLI). Users can generate new dataset instances by specifying parameters such as signal count, sampling resolution, and noise probability directly from the terminal.

290 Technical Validation

291 This section evaluates the signal generation procedure by analyzing spectral properties under different conditions, including the
 292 distribution of dominant frequencies, spectral stability across sampling rates, and the effect of noise. Additionally, we provide a
 293 multi-scale super-resolution benchmark to demonstrate the dataset's utility in training deep learning models, and illustrative
 294 transfer learning experiments using real-world EEG and speech data. These analyses aim to assess variability and stability
 295 under the reported settings, and to document the dataset's behavior for reproducible use. Below, the methodologies and results
 296 are described in detail.

297 Validation Context

298 Experimental parameters were selected to support reproducibility and to illustrate representative behaviors of the generator
 299 under the reported settings. The number of signals ($n = 50$ for spectral analysis, $n = 2500$ for benchmarks) provides a compact
 300 but informative sample to summarize variability. Sampling resolutions (ranging from 150 to 5000 samples) reflect scenarios
 301 requiring different levels of detail, aligning with typical signal processing use cases. Noise amplitudes (Gaussian noise with
 302 $\sigma \in [0.0, 0.2]$) were motivated by common acquisition artifacts, with the goal of providing a controllable benchmark rather
 303 than an exhaustive model of any specific measurement pipeline. Unless otherwise stated, signal-generation settings follow the
 304 configuration in Table 3.

305 Analysis of Dominant Frequency Distribution

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

310 The PSD was estimated using Welch's method, selected for its ability to reduce noise and provide a smoother spectral
 311 representation²². This method stabilizes spectral estimation by dividing the signal into overlapping segments, computing
 312 their individual spectra, and averaging them. This reduces variance from random fluctuations and yields a smoother estimate.
 313 For each signal, the dominant frequency was identified as the frequency at which the PSD reaches its maximum value. This
 314 corresponds to the most prominent spectral component, indicating where the signal concentrates most of its energy.

315 By analyzing the distribution of dominant frequencies across the dataset, we evaluate whether the generated signals exhibit
 316 consistent spectral patterns or if there is significant variation. High consistency would indicate stability in the data generation
 317 process, whereas high variability could suggest the influence of stochastic sampling effects (seed-controlled).

318 The results, shown in Figure 9 and Table 4, indicate that the dominant frequency (reported in Hz under the illustrative convention

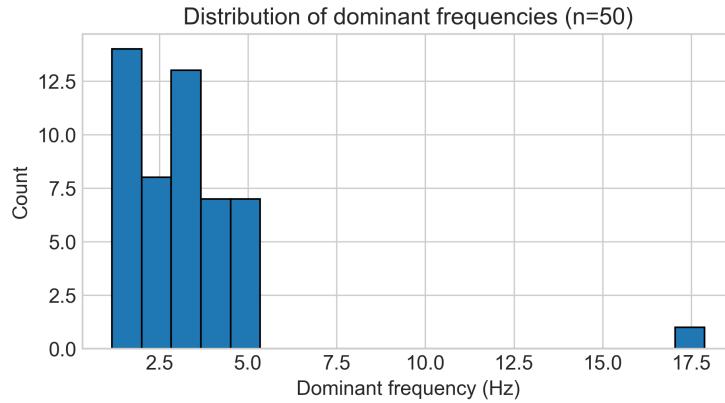


Figure 9. Distribution of dominant frequencies in 50 independently generated signals (reported in Hz, assuming the illustrative convention $T = 4\pi$ s; for other time domains, the axis rescales by $4\pi/T$).

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.

319 $T = 4\pi$ s) is predominantly low under the reported settings. Quantitatively, the mean dominant frequency is 0.508 Hz with a
 320 standard deviation of 0.195 Hz, while the observed range spans 0.390–1.171 Hz (Table 4). Figure 9 shows a strong concentration
 321 in the low-frequency region, with a thinner tail toward higher values, corresponding to a smaller subset of realizations where
 322 the strongest spectral peak shifts upward. This pattern is consistent with the stochastic sampling of frequency profiles (e.g.,
 323 random control points and segment-wise variations): signals share the same parameter ranges, but some draws produce higher
 324 instantaneous-frequency segments that become dominant in the PSD. Overall, the concentration around low frequencies
 325 combined with controlled spread supports the goal of stable primary structure with deliberate diversity, which is desirable for
 326 training models that must generalize across plausible spectral configurations²³.

327 Figure 10 presents examples of signals from the CoSiBD dataset with increasing levels of added noise, illustrating how
 328 amplitude fluctuations progressively obscure the underlying temporal structure.

329 Spectral Stability Across Sampling Resolutions

330 This analysis aims to investigate the influence of sampling resolution (number of samples) on the robustness of spectral estimates
 331 under varying frequency content. When frequency axes are reported in Hz, they follow the illustrative convention $T = 4\pi$ s; for
 332 other choices of T , the Hz axis rescales by $4\pi/T$. At lower resolutions, reduced sampling density and coarser frequency grids
 333 can obscure or merge spectral peaks, compromising the ability to distinguish closely spaced spectral components²⁴. Conversely,
 334 higher resolutions improve the granularity of the frequency axis, allowing for better separation of spectral features and reducing
 335 the risk of misrepresenting the signal's underlying structure²⁵.

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

339 As shown in Figure 11, lower sampling resolutions, specifically the blue curve (150 samples) and the orange curve (250
 340 samples), exhibit a noticeable reduction in detail within the higher-frequency range. These lower-resolution curves display
 341 greater fluctuations, consistent with the theoretical effects of subsampling and aliasing²⁶. In contrast, the higher sampling
 342 resolutions (500, 1000 samples) demonstrate a smoother and more stable spectral profile. This analysis confirms that while
 343 lower sampling rates introduce aliasing artifacts, the dataset provides spectral fidelity comparable to theoretical expectations
 344 when sufficient resolution is employed.

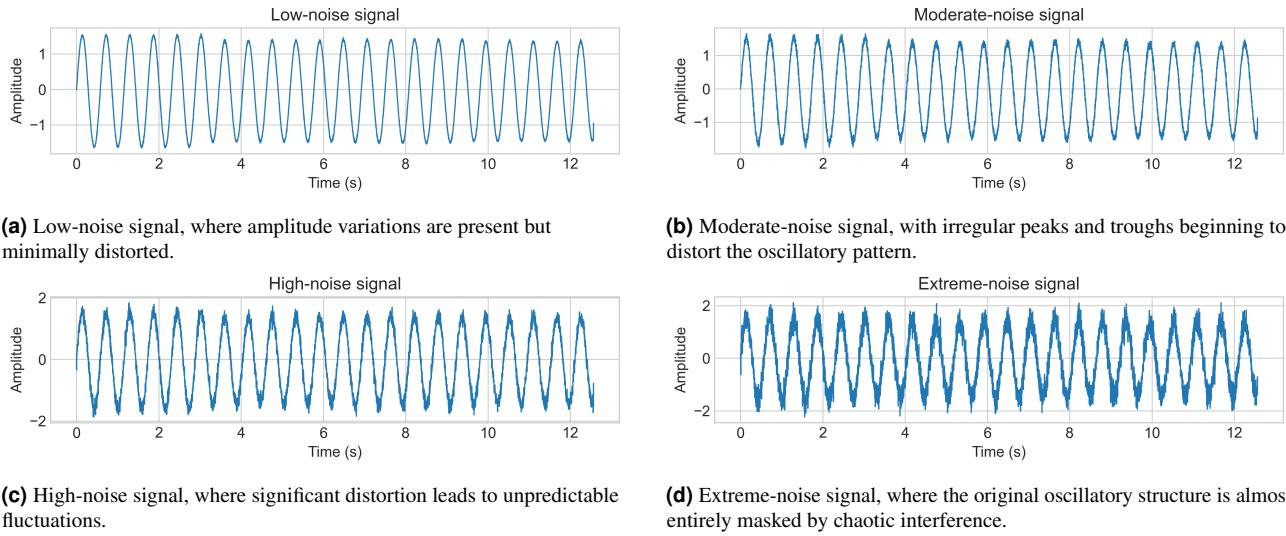


Figure 10. Visualization of signals under increasing noise conditions, showing how added noise progressively masks the original temporal patterns. From low (a) to extreme noise levels (d), this degradation highlights reconstruction challenges for super-resolution models.

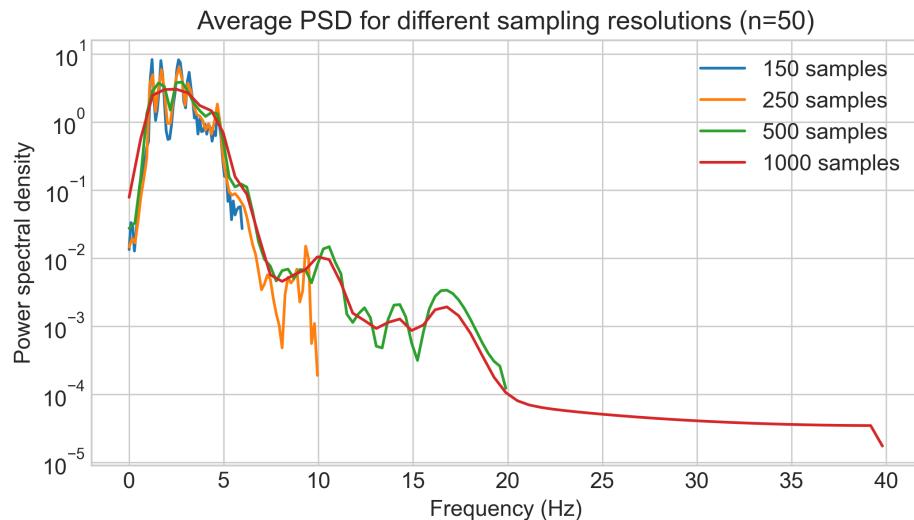


Figure 11. Average power spectral density (PSD) for different sampling resolutions based on 50 independent runs (Hz axis under the illustrative convention $T = 4\pi$ s).

345 **Impact of Noise on Frequency Characteristics**

346 We analyze how varying the noise amplitude affects the power spectral density (PSD), with particular attention to differences
347 between low- and high-frequency regions. Figure 12 illustrates this effect under the reported settings.

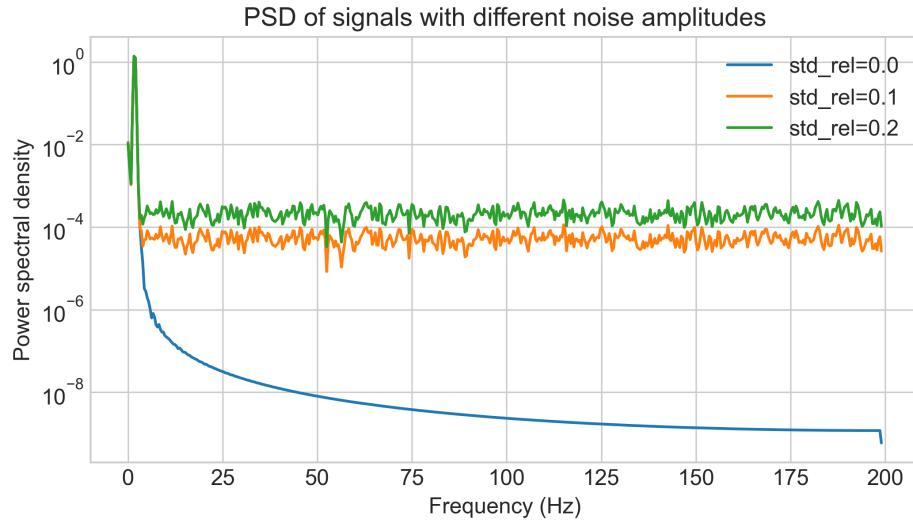


Figure 12. Power spectral density (PSD) of signals generated with different noise amplitudes (Hz axis under the illustrative convention $T = 4\pi$ s).

348 Across these settings, as the noise amplitude increases—from 0.0 (blue curve) to 0.2 (green curve)—the estimated PSD exhibits
349 increased variability at higher frequencies, while the low-frequency region remains comparatively stable. This stability suggests
350 that CoSiBD signals retain their primary structural characteristics even under significant noise, a critical property for robust
351 representation learning²⁷.

352 **Multi-Scale Super-Resolution Benchmark**

353 To illustrate a baseline use case of CoSiBD, we trained a series of convolutional neural network (CNN) models for time
354 series super-resolution at four different scaling factors: $5\times$ ($1000 \rightarrow 5000$), $10\times$ ($500 \rightarrow 5000$), $20\times$ ($250 \rightarrow 5000$), and $33\times$
355 ($150 \rightarrow 5000$). All models employed the TimeSeriesSRNet architecture—a five-layer encoder-decoder network with 1D
356 convolutional layers and bilinear upsampling, inspired by deep residual architectures for audio generation⁹. For this benchmark,
357 the 2,500 high-resolution signals were partitioned into an experiment-specific split of 2,000 paired signals for training and 500
358 held-out signals for validation.

359 We selected this architecture as a simple 1D convolutional encoder-decoder baseline: it captures local temporal structure
360 while providing a deterministic upsampling mechanism, enabling consistent comparisons across scaling factors. Each model
361 was trained using mean squared error (MSE) loss, a standard objective for regression tasks requiring broad mode coverage²⁸,
362 using the Adam optimizer (learning rate 0.001) and early stopping. A batch size of 16 was used as a practical compromise
363 between optimization stability and MPS memory constraints. Training was conducted on Apple Silicon GPU (MPS backend) to
364 accelerate convergence.

365 Table 5 summarizes the validation performance. In these runs, validation loss increased systematically with upsampling factor,
366 reflecting the inherent difficulty of reconstructing fine temporal details from severely undersampled inputs (Table 5, Figure 13).
367 Figure 13 shows the training/validation loss trajectories separately for each scaling factor: (a) $5\times$ ($1000 \rightarrow 5000$), (b) $10\times$
368 ($500 \rightarrow 5000$), (c) $20\times$ ($250 \rightarrow 5000$), and (d) $33\times$ ($150 \rightarrow 5000$). In all four cases, training and validation curves follow similar
369 trends without pronounced divergence, indicating stable optimization under the fixed protocol.

370 Across panels (a–d), the final validation loss increases as the factor grows, consistent with the expected increase in ill-posedness
371 as fewer LR samples constrain the HR target. The $5\times$ and $10\times$ settings (a–b) exhibit faster convergence and lower final error,
372 while the $20\times$ and especially the $33\times$ setting (c–d) show higher residual error, reflecting the greater difficulty of reconstructing
373 fine-scale temporal detail from extreme undersampling.

374 To complement amplitude-based validation, we computed spectral fidelity metrics. Log Spectral Distance (LSD) increased
375 from 0.51 ($5\times$) to 1.21 ($33\times$), indicating progressively larger deviations in spectral magnitude as reconstruction becomes

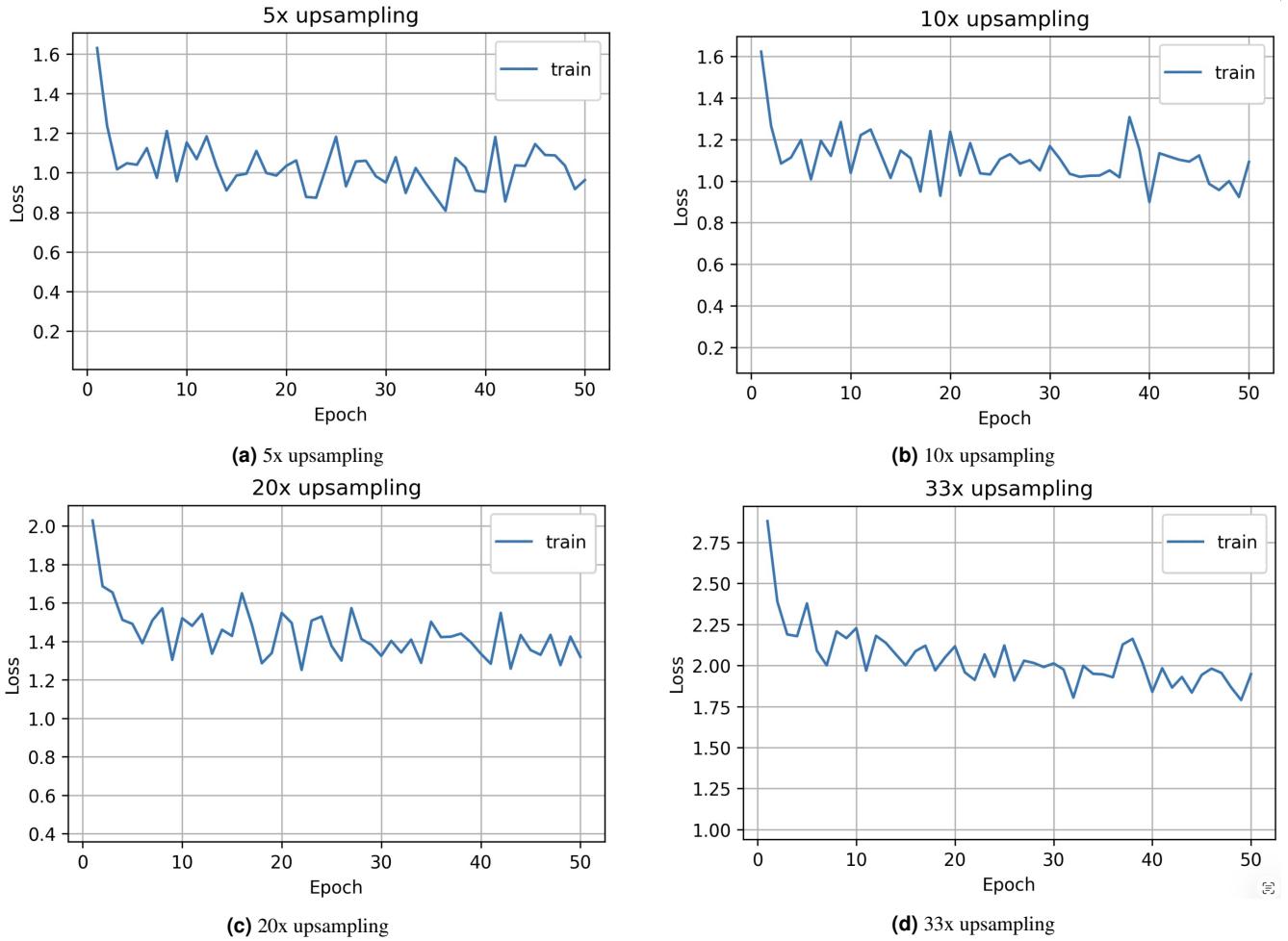


Figure 13. Training/validation loss curves for each upsampling factor in the multi-scale benchmark.

Input Size	Factor	Val Loss	Epochs	Early Stop	LSD	SCORR
1000 samples	5×	0.0845	50	No	0.51±0.63	0.98±0.10
500 samples	10×	0.1524	50	No	0.64±0.63	0.98±0.10
250 samples	20×	0.4376	50	No	0.95±0.67	0.98±0.10
150 samples	33×	1.0326	50	No	1.21±0.67	0.98±0.11

Table 5. Multi-scale super-resolution benchmark results. Validation loss measured as mean squared error on 500 independent validation signals. LSD (Log Spectral Distance) quantifies spectral content deviation, while SCORR (Spectral Correlation) measures frequency-domain similarity. All models completed the full 50-epoch training without early termination, showing stable convergence.

more challenging, while Spectral Correlation (SCORR) remained consistently high (Table 5, Figure 14), suggesting that broad spectral structure is still largely preserved across factors²⁵. Figure 15 provides representative spectrogram comparisons across all upsampling factors; the rightmost column (spectral difference) highlights where reconstruction introduces localized discrepancies (warm colors) that become more pronounced at higher factors, typically around rapid transitions and high-frequency content.

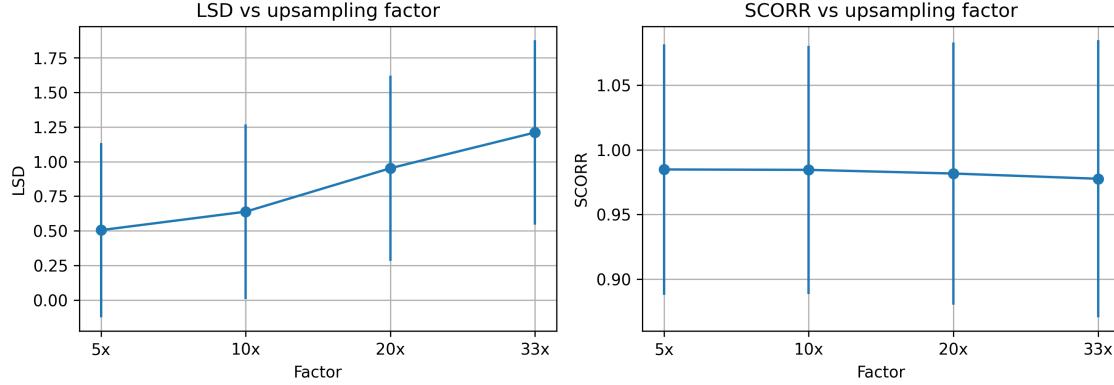


Figure 14. Spectral quality metrics vs upsampling factor. Left: Log Spectral Distance (LSD) increases systematically with upsampling factor, from 0.51 ($5\times$) to 1.21 ($33\times$). Right: Spectral Correlation (SCORR) maintains consistently high values (>0.97) across all factors. Error bars represent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

Representative prediction examples (Figure 16) provide qualitative comparisons of reconstructed outputs against ground truth across scaling factors. Panels (a–b) correspond to $5\times$ upsampling ($1000\rightarrow5000$), (c–d) to $10\times$ ($500\rightarrow5000$), (e–f) to $20\times$ ($250\rightarrow5000$), and (g–h) to $33\times$ ($150\rightarrow5000$). Across all factors, the CNN recovers the main waveform structure and overall trend, indicating that the learned mapping is able to exploit consistent LR–HR correlations in the synthetic pairs.

As the upsampling factor increases, the reconstruction problem becomes increasingly ill-posed: fewer observed samples constrain the HR target, and fine temporal details are progressively harder to recover. This is most apparent at $33\times$, where predictions tend to be smoother and may miss rapid oscillations; correspondingly, a net loss of information at higher frequencies is observed, consistent with the stronger low-pass effect induced by extreme undersampling.

These multi-scale experiments provide quantitative baseline results for future benchmarking studies. In addition to providing a reproducible baseline under a fixed architecture and protocol, the systematic increase in task difficulty—from moderate $5\times$ upsampling to extreme $33\times$ reconstruction—can serve as a practical reference for comparing architectures, loss functions, and training strategies in the time series super-resolution domain. More broadly, CoSiBD enables controlled benchmarking where methods can be compared on identical LR–HR pairs and nuisance settings, supporting fair ablation studies and future community protocols.

Illustrative Transfer Learning Experiments

To demonstrate the practical utility of CoSiBD for training deep learning models, we conducted transfer learning experiments using convolutional neural networks (CNNs) for time-series super-resolution^{9,29}. A TimeSeriesSRNet model (encoder-decoder architecture with 1D convolutions: $1\rightarrow64\rightarrow128\rightarrow256$, bilinear upsampling, decoder $256\rightarrow128\rightarrow64\rightarrow1$) was evaluated on two distinct real-world domains: EEG clinical signals¹ (500 training, 690 validation samples) and VCTK speech recordings³⁰ (44 hours from 109 speakers).

Four training strategies were systematically evaluated: (1) **Real-only**: trained exclusively on domain-specific real data (baseline); (2) **Synth-only**: trained exclusively on CoSiBD synthetic signals; (3) **Mixed**: trained on a combined dataset of synthetic and real data; and (4) **Tuned**: pre-trained on CoSiBD synthetic data, then fine-tuned on the real-world training set. Performance was measured using Mean Absolute Error (MAE) between predicted and ground-truth high-resolution signals.

The results (Table 6, Figure 17) demonstrate that integrating CoSiBD improves model performance compared to using limited real data alone. While models trained only on synthetic data (Synth-only) yielded higher errors due to domain shift, the **Mixed** and **Tuned** strategies consistently outperformed the Real-only baseline. In particular, the best EEG result is an MAE of **9.73** achieved by the **Mixed** strategy, while the best VCTK result is an MAE of **4.41** achieved by **Tuned** (Table 6). Specifically,

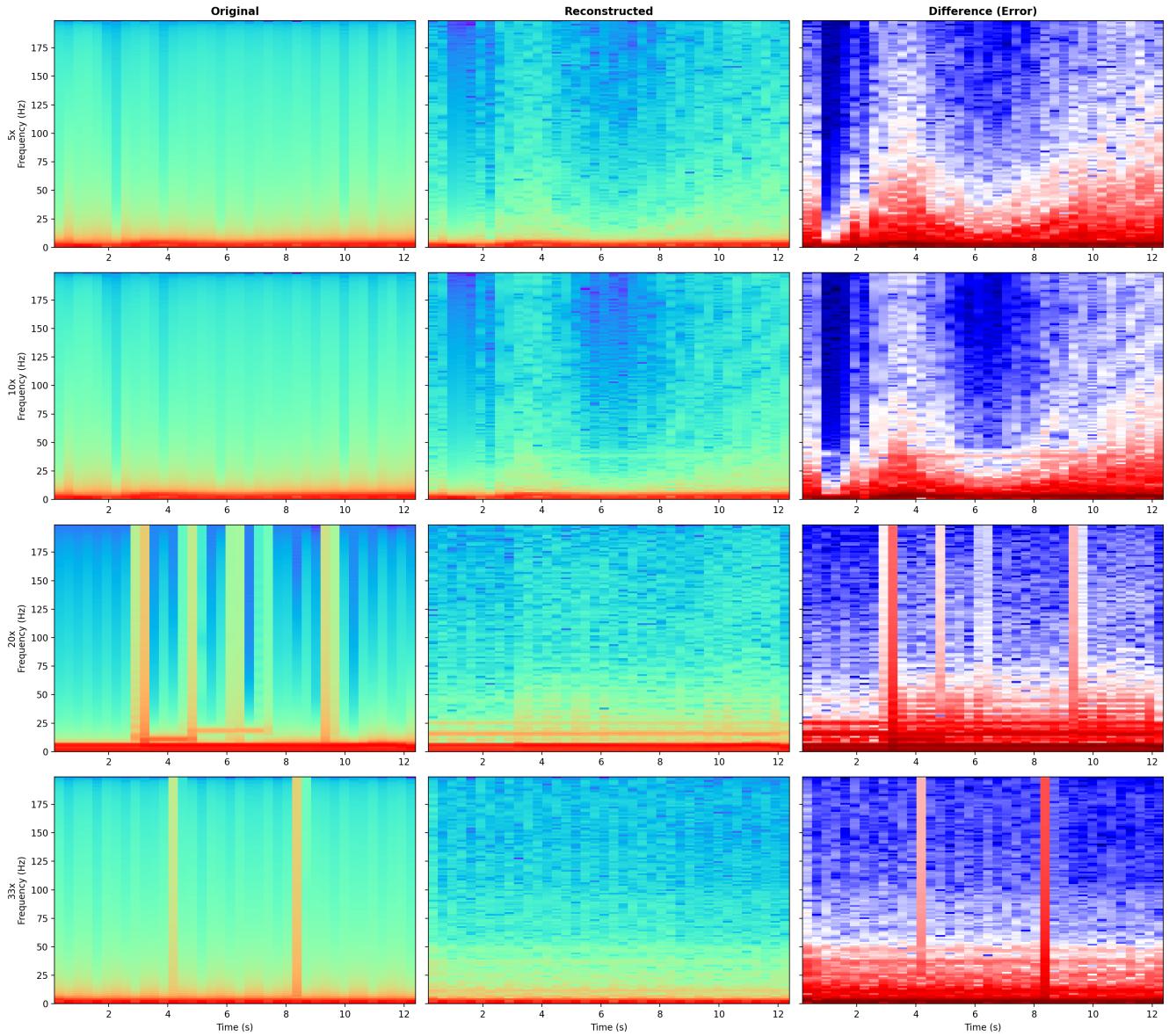


Figure 15. Spectrogram comparison across all upsampling factors. Each row represents a different upsampling factor ($5\times$, $10\times$, $20\times$, $33\times$), showing original signal (left), CNN-reconstructed signal (center), and spectral difference (right). Reconstruction artifacts become more visible at higher upsampling rates. Representative signals selected based on median Log Spectral Distance (LSD) for each factor.

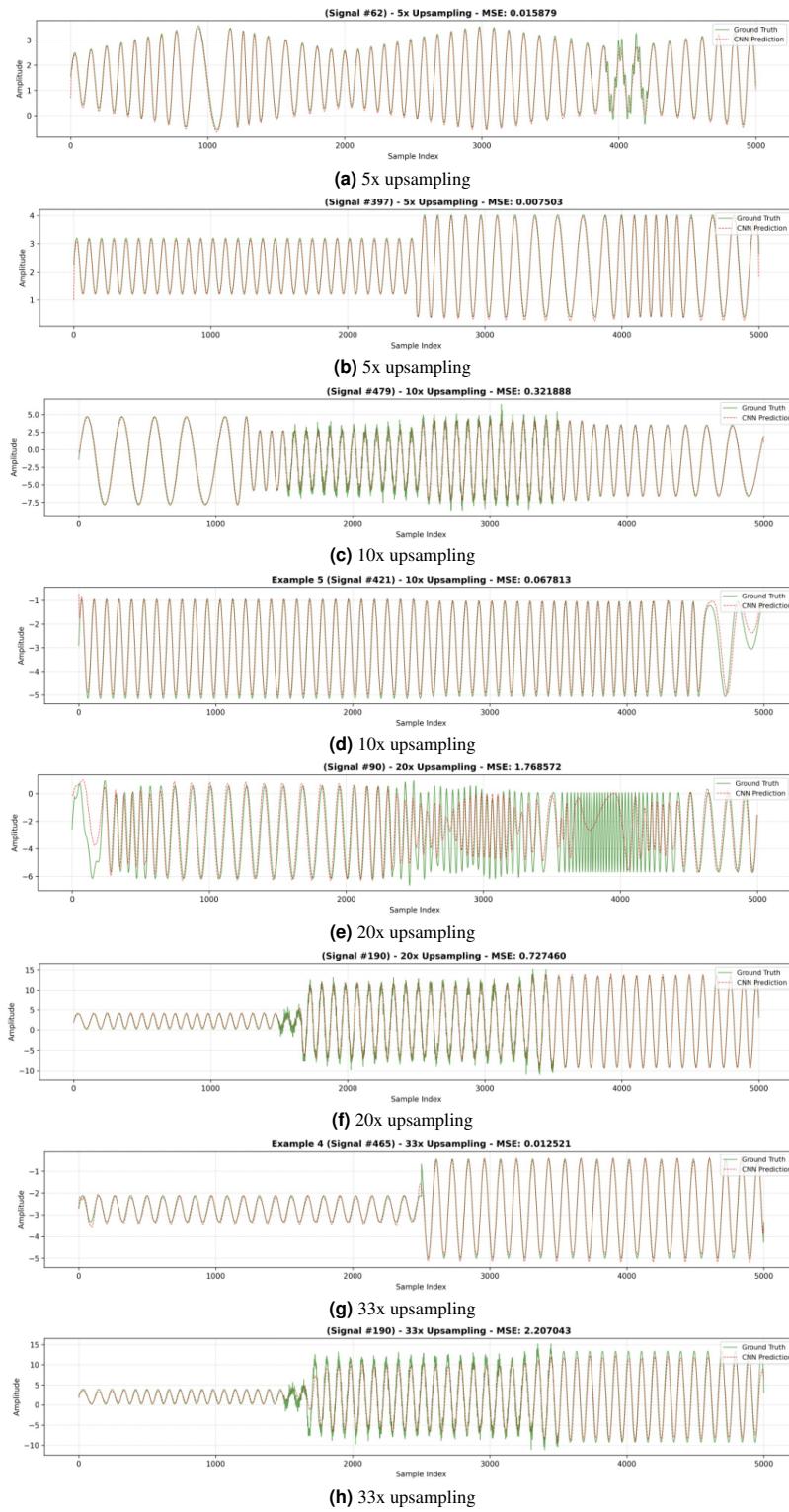


Figure 16. Representative qualitative prediction examples, comparing LR inputs, HR ground truth, and CNN-reconstructed outputs across the multi-scale benchmark. Each panel shows a full-width example.

410 fine-tuning a CoSiBD-pretrained model reduced error significantly on the VCTK dataset, suggesting that the synthetic dataset
411 effectively captures universal temporal structures relevant to super-resolution tasks.

412 Beyond the aggregate MAE values, Figure 17 provides a qualitative interpretation of these results. In the EEG example,
413 Mixed/Tuned reconstructions better follow fast transients and reduce amplitude under/overshoot compared to the baseline. In
414 the speech example, CoSiBD-enhanced strategies recover sharper oscillatory detail and reduce residual smoothing, visually
415 aligning with the lower MAE reported in Table 6.

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)	9.73	5.59
Tuned (pretrain + finetune)	10.68	4.41

Table 6. Mean Absolute Error (MAE) for CNN-based super-resolution models. Embedding CoSiBD data (Mixed and Tuned strategies) improves reconstruction accuracy compared to training on limited real data alone.

416 All those findings suggests that synthetic signals can complement domain-specific realdata. These results are provided as
417 an example of how CoSiBD can be used and depend on the chosen datasets, splits, and training details; they should not be
418 interpreted as definitive claims about general performance but at least in two use cases they prove that our dataset contributes
419 to obtain better results than those obtained using only the real data. This constitutes one of the main contributions of our
420 dataset, although experiments on additional domains are suggested as future research. Detailed experimental methodology
421 and additional comparisons are available in the accompanying repository (see Section of Code availability a the end of the
422 manuscript).

423 As a further validation of structural transferability, with the purpose that the reader can interact with real reconstructed
424 signals, we applied the TimeSeriesSRNet trained solely on CoSiBD ($5\times$ upsampling) to reconstruct full 2-second audio
425 clips from the VCTK corpus without any fine-tuning. Despite the domain gap, the model successfully recovered intelligible
426 speech with preserved harmonic structure (mean Pearson correlation $r = 0.928$). This indicates that CoSiBD's diverse
427 frequency and envelope parameters generalize well to complex, non-stationary signals like human speech. Reconstructed
428 examples (degraded vs. reconstructed signals under different strategies) are provided in the accompanying repository under
429 `AudioModels/results/audio_samples/`.

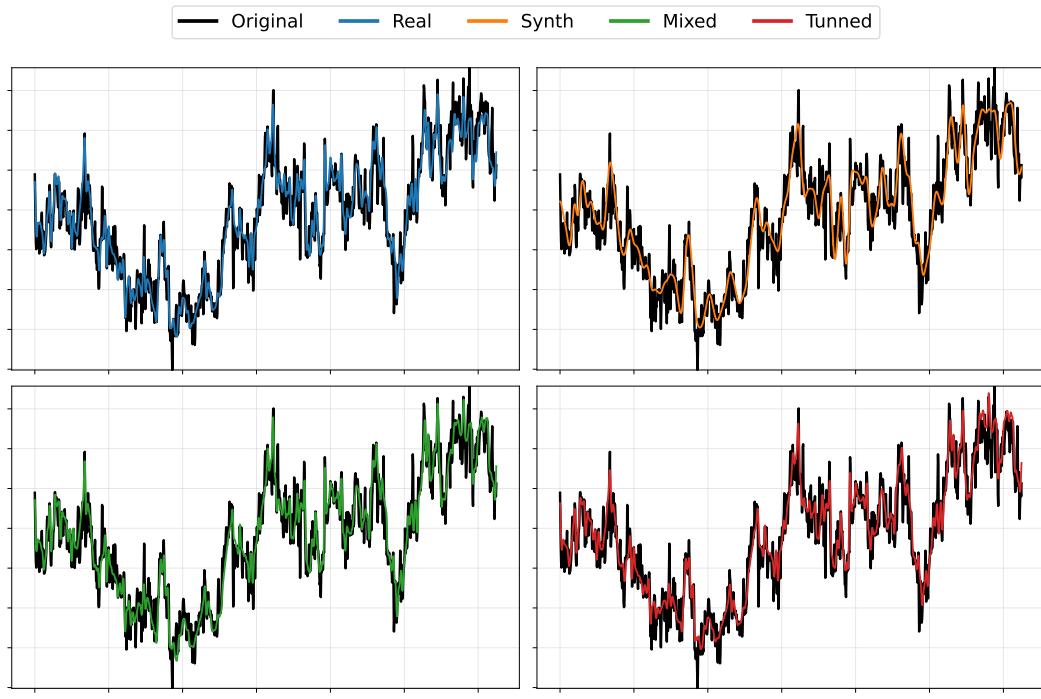
430 Usage Notes

431 The CoSiBD dataset contains high-resolution signals and corresponding subsampled versions at multiple resolutions. Signals
432 are provided in consolidated `.txt`, `.npz`, and `.json` formats. Pairing between low- and high-resolution versions is
433 performed by row index: row i in a subsampled file corresponds to row i in the high-resolution file, with per-signal parameters
434 available in `signals_metadata.json`. The dataset is distributed as a single, unified collection without a predefined
435 train/validation/test split. Users should create partitions appropriate to their objectives (e.g., random splits, stratified splits by
436 noise type/level or signal characteristics, cross-validation, or scenario-specific test sets), using the provided metadata to support
437 principled partitioning.

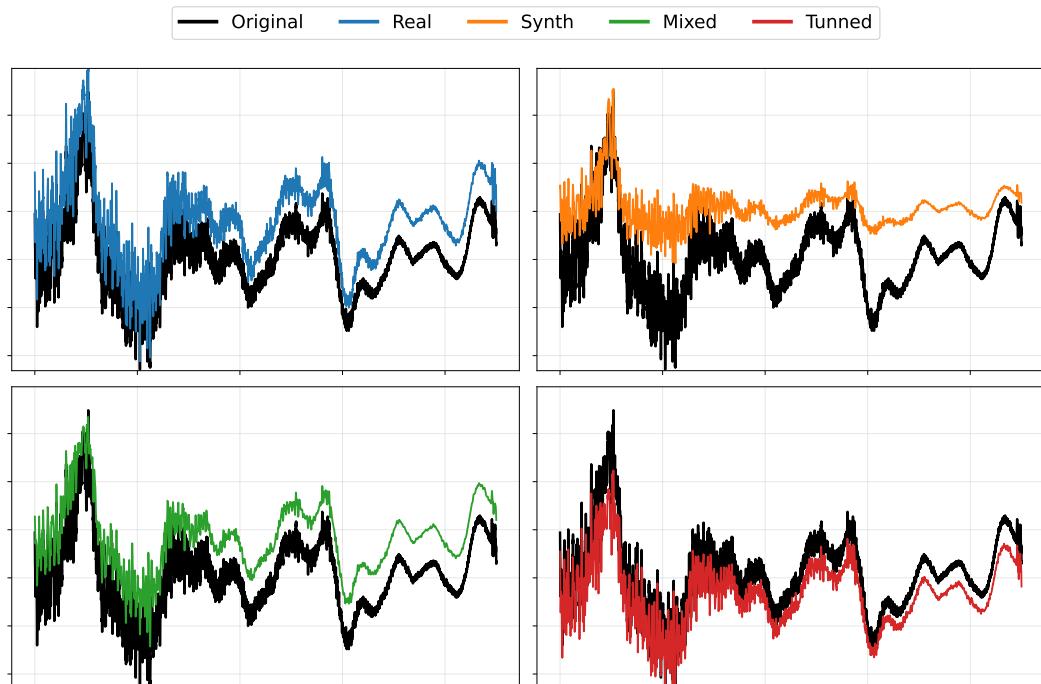
438 Reading the Data

439 In this subsection we show how to read signals stored as consolidated plain text (`.txt`) files, with one signal per row (samples
440 separated by whitespace). Each file contains multiple time series stacked vertically, where each row corresponds to a single
441 signal. The dataset can be accessed using standard Python tools:

```
442 import numpy as np
443
444 # Load subsampled (simple decimation) and high-resolution signals
445 # Each .txt file is consolidated: one signal per row
446 x_valid = np.loadtxt('SignalBuilderC/data/signals_subsampled_simple_250.txt')
447 y_valid = np.loadtxt('SignalBuilderC/data/signals_high_resolution_5000.txt')
448
449 # Optional: convert to PyTorch tensors
450 # import torch
```



(a) EEG clinical signal reconstruction.(x =samples, y=Amplitude)



(b) VCTK speech signal reconstruction. (x =samples, y=Amplitude)

Figure 17. Visual comparison of super-resolution predictions. CoSiBD-enhanced models (Mixed/Tuned) recover finer details in both (a) EEG transients and (b) speech waveforms compared to the baseline.

```
451 # x_valid = torch.tensor(x_valid, dtype=torch.float32)
452 # y_valid = torch.tensor(y_valid, dtype=torch.float32)
453 These commands return NumPy arrays (each row corresponds to one signal). Users can optionally convert them to PyTorch
454 tensors.
```

455 Visualizing Signal Pairs

456 To explore the resolution differences, users can visualize aligned pairs of signals, as indicated in the following example, where
457 well-known functions provided in the Python matplotlib.pyplot interface have been used for that purpose.

```
458 import matplotlib.pyplot as plt
459 import numpy as np
460
461 # Load subsampled (simple decimation) and high-resolution signals
462 # Each .txt file is consolidated: one signal per row
463 x_lr = np.loadtxt('SignalBuilderC/data/signals_subsampled_simple_250.txt')
464 x_hr = np.loadtxt('SignalBuilderC/data/signals_high_resolution_5000.txt')
465
466 # Access a paired signal by row index
467 i = 0
468 low_res_signal = x_lr[i]
469 high_res_signal = x_hr[i]
470
471 # Visualize a paired low- and high-resolution signal
472 plt.figure(figsize=(10, 4))
473
474 # High-resolution signal
475 plt.plot(high_res_signal, label='High-resolution (5000 samples)', alpha=0.8)
476
477 # Low-resolution signal (aligned to HR index range for visualization)
478 lr_x = np.linspace(0, len(high_res_signal), len(low_res_signal))
479 plt.scatter(lr_x, low_res_signal, color='red', s=12,
480             label='Low-resolution (250 samples)')
481
482 plt.xlabel('Sample index')
483 plt.ylabel('Amplitude')
484 plt.title('Paired Low- and High-Resolution Signal')
485 plt.legend()
486 plt.grid(True)
487 plt.tight_layout()
488 plt.show()
```

489 This visualization highlights how the same underlying temporal structure is represented at different resolutions while pre-
490 serving alignment between paired signals. Additional signal characteristics (e.g., change-points, frequency profiles, or noise
491 configuration) can be retrieved from `signals_metadata.json` using the same row index.

492 Training a baseline model (synthetic-only)

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

```
496 import numpy as np
497 import torch
498 import torch.nn as nn
499 from torch.utils.data import DataLoader, TensorDataset
500
501 # Load paired signals (rows align by index)
502 x = np.loadtxt('SignalBuilderC/data/signals_subsampled_simple_250.txt')    # (2500, 250)
```

```

503 y = np.loadtxt('SignalBuilderC/data/signals_high_resolution_5000.txt')    # (2500, 5000)
504
505 # Train/val split (example protocol)
506 x_train, y_train = x[:2000], y[:2000]
507 x_val, y_val = x[2000:2500], y[2000:2500]
508
509 device = torch.device('mps' if torch.backends.mps.is_available() else 'cpu')
510
511 def to_tensor(a):
512     # Convert NumPy array (B, L) into torch tensor (B, 1, L)
513     # The extra channel dimension matches Conv1d input format
514     return torch.tensor(a, dtype=torch.float32).unsqueeze(1)    # (B, 1, L)
515
516 # Create dataloaders for batched training
517 train_loader = DataLoader(TensorDataset(to_tensor(x_train), to_tensor(y_train)),
518                           batch_size=16, shuffle=True)
519 val_loader = DataLoader(TensorDataset(to_tensor(x_val), to_tensor(y_val)),
520                         batch_size=16, shuffle=False)
521
522 class TinySRNet(nn.Module):
523     def __init__(self, out_len=5000):
524         super().__init__()
525         # Encoder: extract local features in the LR domain
526         self.enc = nn.Sequential(
527             nn.Conv1d(1, 64, kernel_size=5, padding=2), nn.ReLU(),
528             nn.Conv1d(64, 128, kernel_size=5, padding=2), nn.ReLU(),
529             nn.Conv1d(128, 256, kernel_size=5, padding=2), nn.ReLU(),
530         )
531         # Upsampling: map LR features to the HR length
532         self.up = nn.Upsample(size=out_len, mode='linear', align_corners=False)
533         # Decoder: project features back to a 1-channel HR signal
534         self.dec = nn.Sequential(
535             nn.Conv1d(256, 128, kernel_size=5, padding=2), nn.ReLU(),
536             nn.Conv1d(128, 64, kernel_size=5, padding=2), nn.ReLU(),
537             nn.Conv1d(64, 1, kernel_size=5, padding=2),
538         )
539
540     def forward(self, x):
541         z = self.enc(x)
542         z = self.up(z)
543         return self.dec(z)
544
545 model = TinySRNet(out_len=5000).to(device)
546 # Optimizer and loss (MSE is a standard regression objective)
547 opt = torch.optim.Adam(model.parameters(), lr=1e-3, weight_decay=1e-5)
548 loss_fn = nn.MSELoss()
549
550 for epoch in range(1, 11):
551     model.train()
552     for xb, yb in train_loader:
553         xb, yb = xb.to(device), yb.to(device)
554         # Forward + loss + backward + update
555         opt.zero_grad()
556         pred = model(xb)
557         loss = loss_fn(pred, yb)

```

```

558     loss.backward()
559     opt.step()
560
561     model.eval()
562     with torch.no_grad():
563         # Validation loop: average loss over batches
564         val_loss = 0.0
565         for xb, yb in val_loader:
566             xb, yb = xb.to(device), yb.to(device)
567             val_loss += loss_fn(model(xb), yb).item()
568         val_loss /= len(val_loader)
569     print(f"epoch={epoch:02d} val_mse={val_loss:.4f}")

```

570 Code availability

571 The complete signal generation pipeline, including modules for frequency profile generation, amplitude envelope construction,
 572 spline interpolation, noise application, and data export in multiple formats, is available at: [CoSiBD scripts on GitHub](#).

573 The repository includes SignalBuilderC, a modular Python package with documented functions for: (1) generating high-
 574 resolution signals with configurable parameters, (2) creating subsampled versions via simple decimation (uniform subsampling),
 575 (3) exporting signals in NumPy, text, and JSON formats, and (4) comprehensive metadata generation. All code is provided with
 576 example notebooks demonstrating dataset regeneration and usage. These scripts are distributed under the MIT License.

577 The dataset itself is published separately at: Zenodo²¹ (DOI: [10.5281/zenodo.15138853](https://doi.org/10.5281/zenodo.15138853)). The Zenodo record distributes the
 578 dataset under the Creative Commons Attribution 4.0 International (CC BY 4.0) license.

579 References

- 580 1. Luciw, M. D., Jarocka, E. & Edin, B. B. Multi-channel eeg recordings during 3,936 grasp and lift trials with varying
 581 weight and friction. *Sci. Data* **1**, 140047, [10.1038/sdata.2014.47](https://doi.org/10.1038/sdata.2014.47) (2014).
- 582 2. Nayak, S. K. *et al.* A review of methods and applications for a heart rate variability analysis. *Algorithms* **16**, 433,
 583 [10.3390/a16090433](https://doi.org/10.3390/a16090433) (2023).
- 584 3. Shaffer, F. & Ginsberg, J. P. An overview of heart rate variability metrics and norms. *Front. Public Heal.* **5**, 258,
 585 [10.3389/fpubh.2017.00258](https://doi.org/10.3389/fpubh.2017.00258) (2017).
- 586 4. Chen, S.-W. Non-uniform sampling data converters: A journey to uncharted circuits and systems. In *2022 International
 587 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).
- 588 5. LeCun, Y., Bengio, Y. & Hinton, G. Deep learning. *Nature* **521**, 436–444, [10.1038/nature14539](https://doi.org/10.1038/nature14539) (2015).
- 589 6. 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)
 590 (2014).
- 591 7. Isasa, I. *et al.* Comparative assessment of synthetic time series generation approaches in healthcare: leveraging patient
 592 metadata for accurate data synthesis. *BMC Med. Informatics Decis. Mak.* **24**, Article 27 (2024).
- 593 8. Zhang, C., Bengio, S., Hardt, M., Recht, B. & Vinyals, O. Understanding deep learning requires rethinking generalization.
 594 *arXiv preprint arXiv:1611.03530* [10.48550/arXiv.1611.03530](https://doi.org/10.48550/arXiv.1611.03530) (2016).
- 595 9. Kuleshov, V., Enam, S. Z. & Ermon, S. Audio super resolution using neural networks. *arXiv preprint arXiv:1708.00853*
 596 [10.48550/arXiv.1708.00853](https://doi.org/10.48550/arXiv.1708.00853) (2017).
- 597 10. Ibarra-Fiallo, J. & Lara, J. A. Contextual deep learning approaches for time series reconstruction. In *2024 IEEE
 598 International Conference on Omni-Layer Intelligent Systems, COINS 2024* (Institute of Electrical and Electronics Engineers
 599 Inc., London, United Kingdom, 2024).
- 600 11. Brophy, E., Wang, Z., She, Q. & Ward, T. Generative adversarial networks in time series: A systematic literature review.
 601 *ACM Comput. Surv.* **55**, Article 199, [10.1145/3559540](https://doi.org/10.1145/3559540) (2023).
- 602 12. McSharry, P. E., Clifford, G. D., Tarassenko, L. & Smith, L. A. A dynamical model for generating synthetic electrocardio-
 603 gram signals. *IEEE Transactions on Biomed. Eng.* **50**, 289–294, [10.1109/TBME.2003.808805](https://doi.org/10.1109/TBME.2003.808805) (2003).

- 604 13. O’Shea, T. J. & West, N. Radio machine learning dataset generation with GNU radio. In *Proceedings of the GNU Radio*
605 *Conference*, vol. 1 (2016).
- 606 14. DeepSig. Datasets (including radioml 2016.10a). <https://www.deepsig.ai/datasets/>. Accessed 2026-01-13.
- 607 15. DeepSig. Radioml 2018.01a dataset. <https://www.deepsig.ai/datasets/>. Accessed 2026-01-13.
- 608 16. McSharry, P. & Clifford, G. D. ECGSYN: A realistic ecg waveform generator (physionet). <https://physionet.org/physiotools/ecgsyn/>. Accessed 2026-01-13.
- 610 17. Krol, L. R., Pawlitzki, J., Lotte, F., Gramann, K. & Zander, T. O. Sereega: Simulating event-related eeg activity. *J. Neurosci. Methods* **309**, 13–24, [10.1016/j.jneumeth.2018.08.001](https://doi.org/10.1016/j.jneumeth.2018.08.001) (2018).
- 612 18. Pinceti, A., Sankar, L. & Kosut, O. Generation of synthetic multi-resolution time series load data. arXiv:2107.03547
613 (2021).
- 614 19. Yuan, Z., Jiang, Y., An, Z., Ma, W. & Wang, Y. Seismic resolution improving by a sequential convolutional neural network.
615 *PLOS ONE* **19**, e0304981, [10.1371/journal.pone.0304981](https://doi.org/10.1371/journal.pone.0304981) (2024).
- 616 20. Sörnmo, L. & Laguna, P. *Bioelectrical Signal Processing in Cardiac and Neurological Applications* (Elsevier Academic
617 Press, 2005).
- 618 21. Ibarra-Fiallo, J., Lara, J. A. & Agudelo Moreno, D. Cosibd, [10.5281/zenodo.15138853](https://zenodo.15138853) (2025). Version v1. Dataset.
- 619 22. Welch, P. D. The use of fast fourier transform for the estimation of power spectra: A method based on time averaging
620 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)
621 (1967).
- 622 23. Bengio, Y., Courville, A. & Vincent, P. Representation learning: A review and new perspectives. *IEEE Transactions on
623 Pattern Analysis Mach. Intell.* **35**, 1798–1828, [10.1109/TPAMI.2013.50](https://doi.org/10.1109/TPAMI.2013.50) (2013).
- 624 24. Rabiner, L. R. & Gold, B. *Theory and Application of Digital Signal Processing* (Prentice Hall, 1975).
- 625 25. Marple, S. L., Jr. *Digital Spectral Analysis with Applications* (Prentice Hall, 1987).
- 626 26. Shannon, C. E. & Weaver, W. *The Mathematical Theory of Communication* (University of Illinois Press, 1949).
- 627 27. Bishop, C. M. *Pattern Recognition and Machine Learning* (Springer, 2006).
- 628 28. Goodfellow, I., Bengio, Y. & Courville, A. *Deep Learning* (MIT Press, 2016).
- 629 29. Kaniraja, C. P., M, V. D. & Mishra, D. A deep learning framework for electrocardiogram (ecg) super resolution and
630 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).
- 631 30. Yamagishi, J., Veaux, C. & MacDonald, K. CSTR VCTK Corpus: English multi-speaker corpus for CSTR voice cloning
632 toolkit (version 0.92), [10.7488/ds/2645](https://doi.org/10.7488/ds/2645) (2019).

633 **Acknowledgments**

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

636 **Author Contributions**

637 J. I. F. handled the methodological design for artificial data creation, probabilistic analysis, spline-based variations, noise
638 distributions, and node selection. J. A. L. was responsible for methodology (time series design) and supervision. D. A. M.
639 performed data processing and validation analysis. All of the authors have contributed to writing the manuscript.

640 **Competing Interests**

641 The authors declare no competing interests