BIOSQ: FREQUENCY-AWARE NEURAL COMPRESSION OF BIOSIGNALS

Anonymous authors
Paper under double-blind review

ABSTRACT

Neural codecs have recently emerged as effective compression models for time-series data, offering a compelling alternative to classical algorithms. For the most part, modern neural codecs are designed to compress audio recordings, such as music and speech, with their architectures and quantization strategies targeted to those modalities. Therefore, extending neural codecs to other modalities poses specific challenges that require careful consideration. In this work, we evaluate the suitability of neural compressors to a variety of biosignals: scalp electroencephalogram (EEG), intracranial EEG, and electrocardiogram (ECG). In particular, we find that global fidelity metrics, commonly used to evaluate compression performance, are insufficient to fully characterize the reconstruction of biosignals. Remarkably, our analysis shows that global fidelity can obscure frequency-localized distortions with disproportionate effects on downstream accuracy. Guided by these findings, we systematically revisit the main components of neural compressors, evaluating alternative quantization strategies and encoder–decoder architectures, and combine them with classical information-theoretic insights. On the one hand, we observe that implicit quantization schemes outperform classical vector quantization. In fact, we introduce residual binary spherical quantization (RBSQ), which extends binary spherical quantization with multiple residual stages, and show that it notably improves fidelity. On the other hand, we find that asymmetric encoder-decoder architectures, while improving reconstruction of audio data, underperform on biosignals. Based on our thorough analysis, we introduce BioSQ as the state-of-the-art (SoTA) neural codec for biosignals. We further reduce the bitrate of BioSQ through entropy regularization and arithmetic coding, promoting stable and uniform symbol distributions. BioSQ outperforms SoTA neural biosignal codecs in reconstruction fidelity across 4 physiological datasets and compression ratios from $4\times$ to $128\times$, reducing reconstruction error by up to 81%. These improvements are also reflected in downstream tasks, where BioSQ better preserves performance on seizure detection, motor-imagery classification, and QRS detection by up to 25% as compression becomes more restrictive. Overall, our results show that neural codec design choices developed for audio do not transfer uniformly to physiological signals, and establish BioSQ as an effective approach to general-purpose biosignal compression.

1 INTRODUCTION

Long-term signal recordings are becoming increasingly common across domains ranging from audio to physiological monitoring, generating large volumes of data that must be efficiently stored, transmitted and analyzed. Neural compression has recently emerged as an effective approach for reducing the size of high-dimensional signals. Neural audio codecs such as SoundStream (Zeghidour et al., 2022) and EnCodec (Défossez et al., 2022), as well as neural image compressors (Ballé et al., 2018), achieve high compression ratios while maintaining excellent reconstruction quality. Despite this success, widely used neural codecs still face important limitations when applied to biosignals: learned codebooks can suffer from inefficient utilization, while common downsampling and upsampling operations can introduce spectral distortions.

The first limitation concerns the quantization strategy. Vector quantization with learned codebooks, often implemented through residual vector quantization (RVQ; Barnes et al. 1996), can suffer from

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

inefficient codebook utilization or codebook collapse, substantially reducing the effective capacity of the discrete representation (Kumar et al., 2023; Zhu et al., 2025). This becomes particularly problematic under restrictive compression budgets, where efficiently using the available representation capacity is critical to preserving signal information. The second consideration concerns the encoder–decoder architecture. State-of-the-art (SoTA) neural codecs (Défossez et al., 2022; Kumar et al., 2023; Zeghidour et al., 2021) commonly rely on architectures originally developed for audio, typically using strided convolutions for downsampling and transposed convolutions for upsampling. Whether these architectural choices are also well suited to physiological signals remains unclear. We therefore systematically investigate alternative quantization strategies and encoder–decoder architectures for neural compression.

Beyond the design of the compression pipeline itself, evaluating reconstruction quality presents a separate challenge. Neural compression methods are commonly evaluated using global reconstruction metrics such as PRD or RMSE (Fauvel & Ward, 2014; Sriraam, 2011; Carzaniga et al., 2025). While these metrics provide an overall measure of reconstruction fidelity, they offer limited insight into how reconstruction errors are distributed across different components of the signal. This limitation can be particularly relevant in domains such as physiological signal compression, where subtle temporal and spectral patterns can carry clinically relevant information (Cole & Voytek, 2017). For example, neurological signals measured with EEG and iEEG contain activity across different frequency bands that has been associated with distinct neural processes and pathological conditions (Newson & Thiagarajan, 2019). Since these bands can differ substantially in signal energy, reconstruction errors in lower-energy components may have little impact on global distortion metrics, despite degrading the corresponding spectral information. Consequently, good global reconstruction quality does not necessarily imply that all relevant signal components are faithfully preserved.

Motivated by these challenges, we focus our study on physiological signals, where codec design and fine-grained reconstruction quality must be considered jointly. Although neural compression is considerably more mature in audio, it remains unclear which of its design choices transfer effectively to physiological signals. Recent studies provide evidence in both directions: pretrained neural audio codecs can provide a strong starting point for EEG compression (Kastrati et al., 2025), while standard audio compression methods can fail to preserve physiologically relevant signal characteristics (Galanti et al., 2026). In particular, the main contributions of this work are:

- We analyze the performance of neural codecs on physiological signals using both global and frequency-specific reconstruction metrics. Our pipeline is motivated by clinically-relevant observations about the morphology of the signals, and confirms that global metrics are insufficient. Indeed, we find that **local errors introduce physiologically relevant degradation that is obscured by global reconstruction quality**.
- We thoroughly examine a variety of quantization strategies and encoder–decoder designs to identify the architecture best suited to compress physiological signals. We find a strong advantage of lookup-free over codebook-based quantization and introduce Residual Binary Spherical Quantization (RBSQ), which further improves fidelity through successive residual stages. BioSQ combines RBSQ with entropy regularization and symmetric encoders and decoders. We further improve compression by up to $1.9\times$ by integrating arithmetic coding directly into the quantizer.
- We evaluate BioSQ across EEG, iEEG, and ECG at compression ratios from $4\times$ to $128\times$, reducing reconstruction error by up to 81% compared with SoTA neural biosignal codecs. We establish BioSQ as the first neural codec effective across both brain and cardiac signals.
- We evaluate the impact of BioSQ on real-world downstream tasks such as seizure detection, motor-imagery classification, and QRS detection. We find that BioSQ improves performance by up to 25% relative to SoTA neural biosignal codecs under high compression.

BioSQ increases the capabilities of neural codecs to include a variety of biosignal modalities, while at the same time improving compression ratios, reconstruction fidelity, and downstream accuracy. Moreover, our detailed analysis of relevant frequency bands clarifies the role of morphological characteristics to the quality of the compression. Indeed, these considerations strongly guide BioSQ's design towards a more physiologically faithful compression solution.

2 Related Work

Biosignal neural compression. While neural codecs have transformed audio compression, their application to biomedical signals remains underexplored. Prior lossy compression methods for EEG rely on wavelet transforms (Xu et al., 2015; Nguyen et al., 2017) compressed sensing (Du et al., 2024) and variational autoencoders (Cisotto et al., 2023), achieving moderate compression ratios with limited reconstruction fidelity. BrainCodec (Carzaniga et al., 2025) introduced the first neural codec for EEG and iEEG, adopting a SEANet encoder-decoder architecture with Residual Vector Quantization (RVQ), substantially outperforming all previous approaches. However, both its encoder-decoder architecture and quantization strategy are largely inherited directly from audio codecs, and its evaluation does not extend to frequency-band or waveform-morphology analysis. Moreover, the original model is limited to brain-derived physiological signals.

Quantization in neural compression. Discrete quantization is a key component of modern neural codecs, mapping continuous latent representations into discrete tokens for efficient transmission and storage. Among the different quantization strategies, Residual Vector Quantization (RVQ) (Barnes et al., 1996) is the most widely adopted, used by SoTA neural audio codecs such as SoundStream (Zeghidour et al., 2021) and EnCodec (Défossez et al., 2022). Despite its widespread adoption, RVQ can suffer from inefficient codebook utilization (Kumar et al., 2023) and codebook collapse (Zhu et al., 2025), while its sequential residual quantization may introduce redundancy across stages. Several works address these limitations while retaining a learned codebook. ERVQ (Zheng et al., 2025) reduces inter-stage redundancy through code-balancing and inter-codebook regularization. SNAC (Siuzdak et al., 2024) instead assigns codebooks of different sizes across multiple temporal resolutions. An alternative line of work eliminates learned codebooks altogether. Representative examples include Finite Scalar Quantization (FSQ) (Mentzer et al., 2023), which discretizes each latent dimension independently, and Binary Spherical Quantization (BSQ) (Zhao et al., 2025), which introduces hyperspherical binary representations.

Encoder-decoder architectures. Beyond the choice of quantizer, encoder-decoder design also shapes the compression quality of a codec. Most neural audio codecs process the waveform directly in the time domain, using stacked convolutional blocks with strided downsampling. VOCOS (Siuzdak, 2024), in contrast, replaces the time-domain decoder with a frequency-domain formulation that predicts the signal's short-time Fourier transform and reconstructs the waveform via an inverse transform. This avoids the aliasing artifacts of transposed convolutions while achieving comparable quality at substantially lower computational cost. WavTokenizer (Ji et al., 2025) builds on this decoder design and adds self-attention layers to better capture long-range dependencies. BigCodec (Xin et al., 2024) instead focuses on the convolutional backbone itself, integrating recurrent layers to model temporal dependencies alongside a larger model capacity.

3 BioSQ: Binary-Quantized Neural Biosignal Compressor

Consider a biosignal $X \in \mathbb{R}^{C \times T}$, with C channels and T samples acquired at sampling frequency f_s. Each channel is processed independently and divided into non-overlapping segments. Each segment is mapped by the encoder to a compact latent representation, quantized using our proposed RBSQ and further compressed using arithmetic coding over the resulting binary indices. The quantized representation is then reconstructed by the decoder.

Encoder and decoder. Each signal segment is processed by an encoder composed of an initial 2D convolution followed by N encoder blocks, where N affects the compression ratio by determining the temporal downsampling before quantization. Each block consists of a residual block followed by a strided 2D convolution with kernel size (K,1) and stride (S,1), with $K = 2S$, progressively downsampling the temporal dimension while increasing the number of channels. A final 2D convolution maps the resulting features to a D-dimensional latent representation. The decoder mirrors the encoder architecture, replacing strided convolutions with transposed convolutions to progressively recover the original temporal resolution. We also evaluate alternative encoder–decoder architectures, including frequency-domain and recurrent designs, but find the symmetric convolutional architecture to perform best (see Section 5.2).

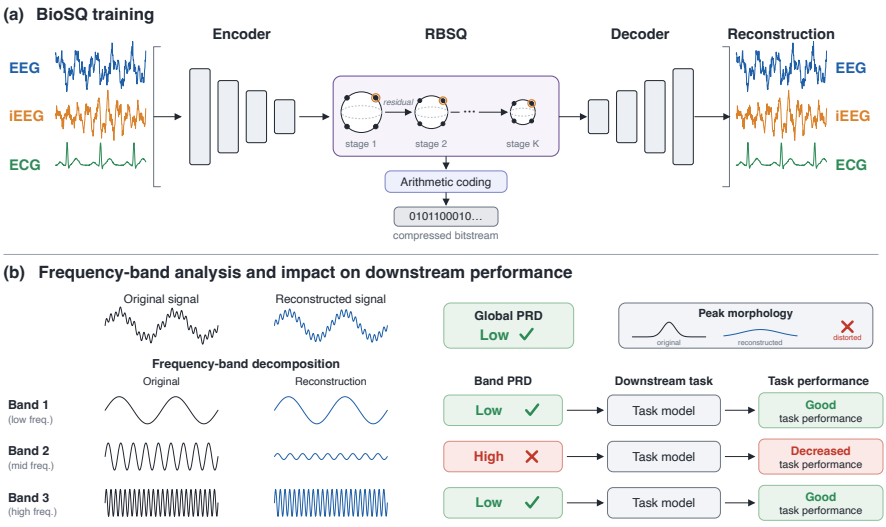


Figure 1: **BioSQ architecture and evaluation.** **(a)** BioSQ compresses EEG, iEEG, and ECG signals using a SEANet encoder and decoder with Residual Binary Spherical Quantization (RBSQ), followed by lossless arithmetic coding of the binary indices. **(b)** Reconstruction fidelity is evaluated both globally and across frequency bands, showing that low global distortion can mask larger errors in specific frequency bands that may affect downstream task performance.

Residual Binary Spherical Quantization (RBSQ). We find that lookup-free quantization outperforms vector quantization in reconstruction fidelity (Section 5.2). We further find that iterative quantization benefits lookup-free methods, motivating the introduction of residual binary spherical quantization (RBSQ). At each stage k, the residual $\mathbf{r}_k \in \mathbb{R}^D$ is projected to d dimensions, L_2-normalized, binarized and projected back to the latent space via independent per-stage projections $W_k^{\text{in}} \in \mathbb{R}^{d\times D}$ and $W_k^{\text{out}} \in \mathbb{R}^{D\times d}$:

$$\tilde{\mathbf{z}}_k = \frac{W_k^{\text{in}}\mathbf{r}_k}{||W_k^{\text{in}}\mathbf{r}_k||_2}, \qquad \mathbf{q}_k = W_k^{\text{out}}\left(\text{sign}(\tilde{\mathbf{z}}_k)\cdot\frac{1}{\sqrt{d}}\right), \qquad \mathbf{r}_{k+1} = \mathbf{r}_k - \mathbf{q}_k \tag{1}$$

After K stages, the final quantized representation is obtained as $\hat{\mathbf{z}} = \sum_{k=1}^{K}\mathbf{q}_k$. Each stage uses independent input and output projections, initialized orthogonally. Since the sign operation is non-differentiable, we use a straight-through estimator during training.

Entropy regularization. We observe that RBSQ still suffers from inefficient utilization despite the lack of an explicit codebook. Following BSQ (Zhao et al., 2025), we adopt entropy regularization to encourage confident and balanced binary assignments. We extend this approach to RBSQ by regularizing each residual stage $k \in \{1,\dots,K\}$ independently, as successive stages operate on different residual distributions. For each component $v_{k,i}$ of the projected representation $\mathbf{v}_k = W_k^{\text{in}}\mathbf{r}_k$, we define a soft binary probability $p_{k,i} = \sigma(v_{k,i})$, where σ denotes the sigmoid function. We therefore minimize the binary entropy while additionally encouraging balanced bit usage:

$$H(p_{k,i}) = -p_{k,i}\log p_{k,i} - (1-p_{k,i})\log(1-p_{k,i}), \qquad \mathcal{L}_{\text{bal}}^{(k)} = \frac{1}{d}\sum_{i=1}^{d}\left(\bar{p}_{k,i} - \frac{1}{2}\right)^2 \tag{2}$$

where $\bar{p}_{k,i}$ denotes the mean probability of bit i over the batch. The entropy term encourages confident assignments, while the balance term prevents individual bits from collapsing to a single value. The final loss (see Appendix H) combines both terms, averaged across the K residual stages and weighted by λ_{ent}.

Arithmetic coding. The $K \times d$ binary indices produced per frame are not uniformly distributed, particularly at early stages where probability mass concentrates on a subset of the 2^d possible symbols. We exploit this redundancy by applying a static arithmetic coder per stage, treating each stage's d bits as a single symbol over a 2^d-symbol alphabet and estimating its empirical distribution from a held-out set. The effective compression ratio after arithmetic coding is: $\mathrm{CR}_{\mathrm{eff}} = \mathrm{CR} \times (K \cdot d) \,/\, (\sum_k H_k)$, where H_k is the empirical Shannon entropy of stage k in bits. This operation is lossless: the decoder receives the same indices, leaving PRD unchanged.

4 Training Setup

We process each channel independently using non-overlapping 4-second segments. For the encoder and decoder, we set $F = 16$, $S = 2$ and $D = 64$, with the number of channels doubled after each downsampling layer up to a maximum of 256, and use ELU activations throughout. For RBSQ, we use $K = 8$ residual stages with $d = 4$ binary dimensions per stage, yielding a total budget of $K \times d = 32$ bits per latent frame (ablation in Appendix F). We set the entropy regularization weight to $\lambda_{ent} = 0.1$. For arithmetic coding, we use $d = 4$ bits per residual stage, corresponding to an alphabet of $2^d = 16$ possible symbols per stage. BioSQ is trained end to end with $\mathcal{L} = \mathcal{L}_{\text{rec}} + \mathcal{L}_{\text{spec}} + \lambda_{\text{ent}}\mathcal{L}_{\text{ent}}$, combining an ℓ_1 waveform loss, a multi-resolution spectral loss (Appendix H), and the entropy regularization of Eq. 2. Unlike prior neural codecs, we use no adversarial training.

4.1 Datasets

We evaluate BioSQ on four public datasets covering three physiological signal modalities: EEG (CHB-MIT, BCI Competition IV-2a) for seizure detection and motor imagery classification, iEEG (SWEC-ETHZ) for seizure detection, and ECG (MIT-BIH Arrhythmia) for heartbeat detection.

CHB-MIT (Shoeb, 2010). Contains a total of 24 subjects, 983 hours of recording and 198 seizures. All signals are sampled at 256 Hz. Onset and offset times of the seizures are included in the dataset.

BCI Competition IV-2a (Tangermann et al., 2012). Nine subjects performed four motor imagery tasks, with each trial labeled according to the corresponding task. Signals are acquired from 22 EEG channels at 250 Hz and band-pass filtered between 0.5 and 100 Hz, with a 50 Hz notch filter to remove power-line interference.

SWEC-ETHZ iEEG (Burrello et al., 2019). Recordings from 15 subjects, with a total duration of 14 hours and 104 seizure events. Signals are sampled at either 512 Hz or 1024 Hz and preprocessed using median referencing and a 0.5–120 Hz fourth-order Butterworth band-pass filter. Seizure onset and offset times are annotated by an expert neurologist, who excluded channels affected by artifacts.

MIT-BIH Arrhythmia (Goldberger et al., 2000). Contains 48 half-hour two-channel ambulatory recordings collected from 47 subjects for cardiac arrhythmia analysis. Signals are sampled at 360 Hz. Each recording includes expert annotations for approximately 110,000 heartbeats, providing the location and type of each beat.

4.2 Baselines

We compare the performance of BioSQ against multiple lossy compression algorithms developed for EEG: BrainCodec (Carzaniga et al., 2025), a neural biosignal codec based on Residual Vector Quantization (RVQ); NLSPIHT (Xu et al., 2015), a classical algorithm; and AAC (Nguyen et al., 2017), a wavelet-based model with adaptive arithmetic coding.

4.3 Evaluation Metrics

To evaluate the compression performance in terms of reconstruction fidelity, we report the percentage root-mean-square distortion (PRD), and, for EEG and iEEG signals, a band-specific variant

PRD$_b$ computed after FFT bandpass filtering x and $\hat{x}$ over physiologically relevant frequency bands (δ: 0.5–4 Hz, θ: 4–8 Hz, α: 8–13 Hz, β: 13–30 Hz, γ_{low}: 30–70 Hz, γ_{high}: 70–120 Hz):

$$\text{PRD} = \frac{||x - \hat{x}||_2}{||x||_2} \cdot 100, \qquad \text{PRD}_b = \frac{||x_b - \hat{x}_b||_2}{||x_b||_2} \cdot 100 \tag{3}$$

Downstream tasks. While PRD quantifies signal fidelity, it does not directly reflect whether clinically relevant information is preserved under compression. Following the procedure of BrainCodec (Carzaniga et al., 2025), we assess task performance when inference is performed on compressed signals, training a subject-specific classifier on the original signal and evaluating it on the reconstructed one. We apply this to seizure detection on CHB-MIT and SWEC-ETHZ using EEGWaveNet (Thuwajit et al., 2022), and to motor imagery classification on BCI IV-2a using MIBMInet (Wang et al., 2024). We expand this evaluation to QRS detection on MIT-BIH using the Pan-Tompkins algorithm (Pan & Tompkins, 1985), reporting the fraction of annotated R-peaks recovered within $\pm$75 ms of each ground-truth annotation.

Morphological features. We evaluate whether compression distorts clinically meaningful signal features by applying standard extractors to both original and reconstructed signals, as ictal waveform morphology has been linked to pathological brain states (Cole & Voytek, 2017). Specifically, we measure amplitude error (relative peak amplitude distortion), sharpness error (second-derivative at the peak), and rise-decay asymmetry (difference in slope between the rising and falling flanks) on ictal peaks detected in CHB-MIT and SWEC-ETHZ. The detection procedure and the results are reported in Appendix D.

5 Results

5.1 Frequency-Aware Evaluation

In audio, human perception can fill in small spectral gaps (Painter & Spanias, 2000), such that good global error is sufficient to guarantee the quality of the reconstructed signal. We argue that physiological signals offer reduced tolerance to spectral distortions. Indeed, downstream classification pipelines can be sensitive to errors in specific frequency bands, and human experts can be affected by distorted morphological features as well (Cole & Voytek, 2017). Therefore, we test whether global PRD alone adequately characterizes the downstream impact of reconstruction errors.

We provide the downstream classification model with two different reconstructions of the CHB-MIT dataset at CR=64×: (1) the generalist reconstruction, obtained through the standard BioSQ; (2) the specialist reconstruction, obtained from a combination of 6 BioSQ, each trained on a specific band

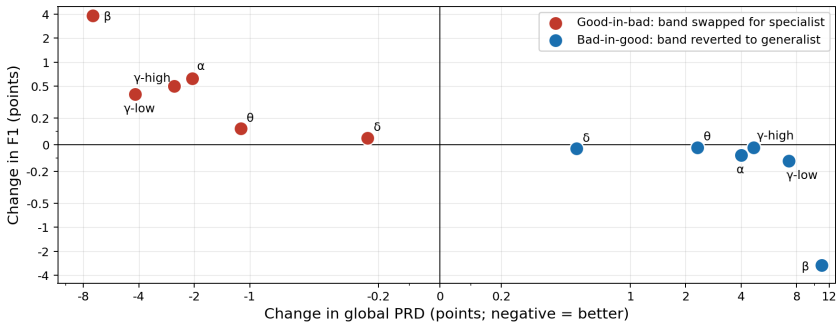


Figure 2: **Per-band reconstruction quality, global PRD, and downstream F1.** CHB-MIT seizure detection at CR=64×. We swap one frequency band's reconstruction between two reconstructions: the generalist (one model trained on all bands) and the specialists (six models, each trained on a single band). Each point plots the resulting change in global PRD against the change in downstream F1, when improving one band (good-in-bad, red) or degrading one band (bad-in-good, blue).

only (see Appendix E). Using these two versions, we build composite datasets by replacing one band at a time from generalist to specialist, and vice-versa.

We evaluate the two symmetric settings. In *good-in-bad*, we replace one band of the generalist reconstruction with the same band from the specialist, improving the reconstruction fidelity of that band. In *bad-in-good*, we replace one band of the specialist reconstruction with the same band from the generalist, degrading the fidelity of that band. Finally, to quantify the discrepancy between global PRD and downstream performance, we measure the effect of each swap on seizure detection.

Figure 2 shows, for each composite representation, the change in global PRD against the change in F1. We find that changes in global PRD do not translate proportionally into changes in accuracy. In particular, perturbations with similar effects on global PRD can have markedly different downstream consequences depending on the affected frequency band. In particular, the β band has a disproportionate effect in both directions ($\Delta F_1 = -3.84$ for *bad-in-good*, $\Delta F_1 = +3.84$ for *good-in-bad*). At the same time, γ_{low} has a similar effect on global PRD, yet its downstream effect is minimal.

5.2 Architectural Exploration

We show in Section 5.1 that global PRD can obscure how reconstruction error is distributed across frequencies. We therefore complement global evaluation with per-band PRD (Figure 3). This includes the quantizer and the encoder-decoder architecture, discussed below, as well as arithmetic coding of the resulting indices, discussed separately in Section 5.3 and Appendix C.

Quantizer. We first compare different quantization strategies. We evaluate RVQ together with ERVQ (Zheng et al., 2025) and SNAC (Siuzdak et al., 2024), which are techniques that aim to improve codebook utilization and reduce redundancy. We find that both provide limited gains over standard RVQ. In contrast, lookup-free methods achieve lower reconstruction error than codebook-based approaches. Among them, extending BSQ (Zhao et al., 2025) with residual quantization yields the best performance, motivating our choice of RBSQ for BioSQ. The top row of Figure 3 shows a representative subset of this comparison (the full quantizer exploration is in Appendix G).

Encoder-Decoder. We next compare different encoder–decoder architectures in the bottom row of Figure 3. We first evaluate a Vocos-based architecture (Siuzdak, 2024), replacing the convolu-

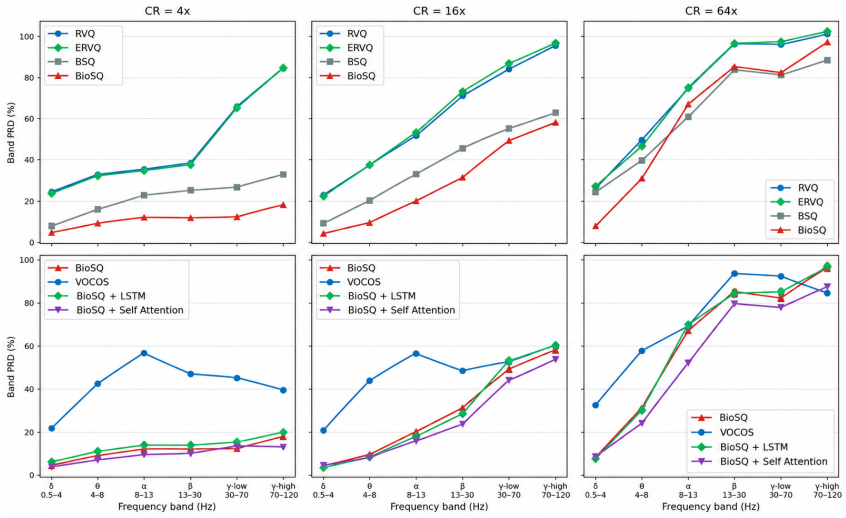


Figure 3: **Band-wise reconstruction error across quantizers and encoder-decoder architectures.** PRD is computed independently over six frequency bands at compression ratios from 4× to 64×. The top row varies the quantizer while keeping the encoder and decoder architecture fixed. The bottom row varies the encoder-decoder architecture while keeping RBSQ, our final quantizer choice, fixed. Lower values indicate better reconstruction fidelity within each frequency band. For clarity, only a subset of quantizer variants is shown here; the full comparison is in Appendix G.

tional downsampling and upsampling operations with frequency-domain transformations to mitigate aliasing artifacts. We then adapt the LSTM layers used in BigCodec (Xin et al., 2024) to the BioSQ architecture to improve temporal modeling. Finally, following WavTokenizer (Ji et al., 2025), we use its decoder, a Vocos backbone with a self-attention block, in BioSQ to enhance contextual modeling. In all cases, we retain RBSQ to isolate the effect of the encoder–decoder architecture. The Vocos-based architecture underperforms across nearly all frequency bands and compression ratios, and the LSTM variant provides no consistent benefit over the baseline. In contrast, we find that insertion of a self-attention block consistently improves reconstruction fidelity, albeit at a significant computational cost of 17× more parameters and 9× reduced throughput at CR=16×. To keep BioSQ lightweight, we do not adopt self-attention for the main model.

5.3 Compression performance

We now evaluate BioSQ against BrainCodec, AAC (Nguyen et al., 2017), and NLSPIHT (Xu et al., 2015) across datasets and compression ratios. Figure 4 shows median PRD across held-out subjects for all datasets and compression ratios. BioSQ achieves lower PRD than all baselines on both EEG datasets, CHB-MIT and BCI IV-2a, at every compression ratio, but the two datasets behave differently as compression increases. On CHB-MIT, the advantage with respect to BrainCodec narrows with compression, whereas on BCI IV-2a it marginally increases. On the iEEG SWEC-ETHZ dataset, BrainCodec and AAC achieve lower PRD at CR=4×, while BioSQ performs better from CR=16× onward. We also evaluate the ECG MIT-BIH dataset, where we extend the baseline BrainCodec to this new modality. Here, BioSQ suffers at lower compression ratios, but outperforms the baseline as compression increases. Interestingly, we show in Section 5.4 that the relatively lower performance of BioSQ at low compression ratios has no noticeable effect on downstream classification, while the improved fidelity at high compression ratios does. Representative waveform and spectral reconstructions for CHB-MIT and SWEC-ETHZ are shown in Appendix B.

Arithmetic coding. Beyond reconstruction fidelity, we test whether BioSQ's discrete representations can be compressed further with lossless arithmetic coding. The gain varies by modality: it stays close to the nominal ratio for EEG and iEEG, but reaches up to 1.9× on ECG (MIT-BIH), where the learned binary representations are more predictable. We investigate this further in Appendix C, where we also quantify the effect of arithmetic coding on BrainCodec.

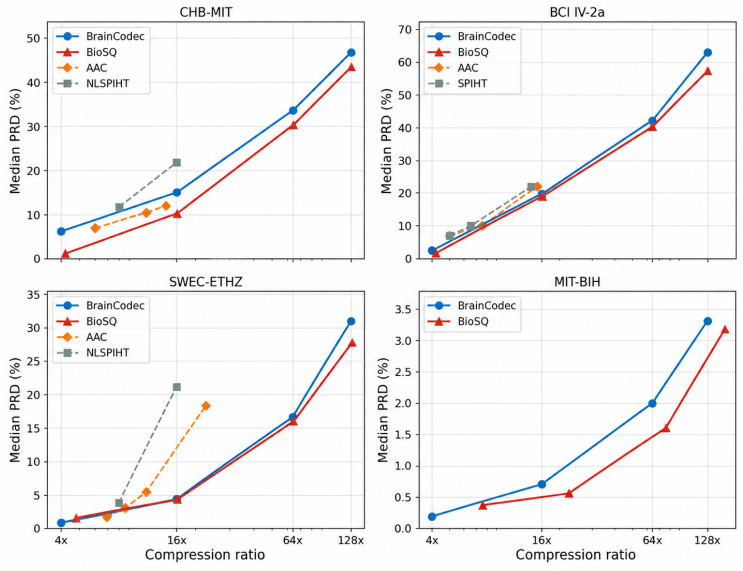


Figure 4: **Median PRD across compression ratios and datasets** Each point is the median PRD over held-out subjects. For BioSQ, the x-axis reports the effective compression ratio after arithmetic coding rather than the nominal one. BioSQ matches or outperforms BrainCodec across most datasets and compression ratios, with a crossover in relative performance emerging on SWEC-ETHZ and MIT-BIH. AAC and NLSPIHT results, where available, are taken from (Carzaniga et al., 2025).

5.4 DOWNSTREAM CLASSIFICATION TASKS

We complete our analysis by verifying the efficacy of the compressors on downstream tasks. Figure 5 reports the performance across all datasets. BioSQ matches or outperforms BrainCodec downstream at every compression ratio on every dataset. On CHB-MIT, the advantage of BioSQ follows from the notably lower PRD with respect to BrainCodec. BioSQ matches the original seizure detection sensitivity at CR=16×, a fairly aggressive compression ratio, while BrainCodec declines. The result is more surprising on BCI IV-2a, SWEC-ETHZ, and MIT-BIH. On BCI IV-2a, the two codecs reconstruct the signal almost equally well, yet BioSQ still wins downstream at every compression ratio. On SWEC-ETHZ and MIT-BIH, BrainCodec actually achieves lower PRD at low compression ratios, but this reconstruction advantage does not translate into a better F1 score (see Appendix A for more details). In contrast, BioSQ achieves notably higher performance at high compression ratios.

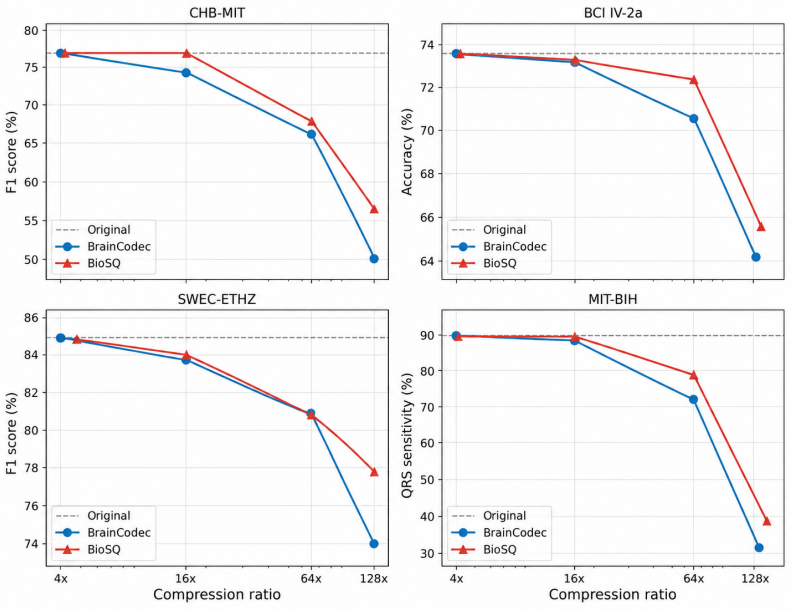


Figure 5: **Downstream task performance of BioSQ and BrainCodec** The reported metrics are motor imagery classification accuracy for BCI IV-2a, seizure detection F1 score for CHB-MIT and SWEC-ETHZ, and QRS detection sensitivity for MIT-BIH.

6 CONCLUSION

We find that global reconstruction metrics alone can obscure frequency-localized errors with disproportionate effects on downstream performance.. To capture these errors, we evaluate neural compression of biosignals with frequency-aware metrics in addition to global ones, and use them to thoroughly compare quantization strategies and encoder-decoder architectures. We find that our residual binary spherical quantization (RBSQ), paired with a symmetric convolutional encoder-decoder, achieves the best performance. Combining the results of our frequency-aware evaluation with entropy regularization and arithmetic coding we obtain BioSQ, our biosignal neural codec. BioSQ is the first neural compressor trained on EEG, iEEG, and also ECG, generalizing this technique beyond brain signals. In particular, BioSQ reduces reconstruction error by up to 81% over the SoTA across 4 datasets and compression ratios ranging from 4× to 128×. Higher reconstruction fidelity also leads to improved downstream performance, with BioSQ outperforming BrainCodec by up to 12% in seizure detection, 2% in motor-imagery classification, and 22% in QRS detection.

Limitations. At the highest compression ratios, high-frequency activity and sharp waveform features remain hard to preserve, and downstream performance degrades. Further work is therefore needed to obtain a more balanced fidelity across frequency bands.

AI USE STATEMENT

In this work, we used generative AI tools to edit and polish the text of the paper, including connecting the main text with the appendices; to write plotting code for several figures; and to produce a draft of the schematic in Figure 1, which we then revised. We did not use generative AI tools to generate the core method or the trained models, which were designed and implemented by the authors. We have reviewed all AI-assisted work: the code was run and its outputs checked by the authors, and the text and figures were revised by the authors. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work uses only publicly available, de-identified recordings (CHB-MIT, BCI Competition IV-2a, SWEC-ETHZ, and MIT-BIH Arrhythmia), following the respective license. No new human data was collected in the context of this work. BioSQ may only be used for research. In particular, this model must not be used for diagnostic purposes.

REPRODUCIBILITY STATEMENT

All datasets used are publicly available (Section 4.1). The architecture, training setup, and loss are described in Section 4 and Appendix H and the RBSQ hyperparameters are ablated in Appendix F. As indicated by ICLR's author guidelines, we are available to share the code anonymously during the rebuttal period.

REFERENCES

Johannes Ballé, David Minnen, Saurabh Singh, Sung Jin Hwang, and Nick Johnston. Variational image compression with a scale hyperprior. *arXiv preprint arXiv:1802.01436*, 2018.

Christopher F. Barnes, Syed A. Rizvi, and Nasser M. Nasrabadi. Advances in residual vector quantization: A review. *IEEE Transactions on Image Processing*, 5(2):226–262, 1996. doi: 10.1109/83.480761.

Alessio Burrello, Lukas Cavigelli, Kaspar Schindler, Luca Benini, and Abbas Rahimi. Laelaps: An energy-efficient seizure detection algorithm from long-term human iEEG recordings without false alarms. In *Proceedings of the Design, Automation & Test in Europe Conference & Exhibition (DATE)*, pp. 752–757. IEEE, 2019. doi: 10.23919/DATE.2019.8715186.

Francesco Stefano Carzaniga, Gary Tom Hoppeler, Michael Hersche, Kaspar Anton Schindler, and Abbas Rahimi. The case for cleaner biosignals: High-fidelity neural compressor enables transfer from cleaner iEEG to noisier EEG. *arXiv preprint arXiv:2502.17462*, 2025.

Giulia Cisotto, Alberto Zancanaro, Italo F. Zoppis, and Sara L. Manzoni. hvEEGNet: Exploiting hierarchical VAEs on EEG data for neuroscience applications. *arXiv preprint arXiv:2312.00799*, 2023.

Scott R. Cole and Bradley Voytek. Brain oscillations and the importance of waveform shape. *Trends in Cognitive Sciences*, 21(2):137–149, 2017. doi: 10.1016/j.tics.2016.12.008.

Alexandre Défossez, Jade Copet, Gabriel Synnaeve, and Yossi Adi. High fidelity neural audio compression. *arXiv preprint arXiv:2210.13438*, 2022.

X. Du, K. Liang, Y. Lv, and S. Qiu. Fast reconstruction of EEG signal compression sensing based on deep learning. *Scientific Reports*, 14(1):5087, 2024. doi: 10.1038/s41598-024-55334-9.

Simon Fauvel and Rabab K. Ward. An energy efficient compressed sensing framework for the compression of electroencephalogram signals. *Sensors*, 14(1):1474–1496, 2014.

Mattia Galanti, Filippo Battaglia, Giovanni Gugliandolo, Petter Omland, Kristian Nilsen, Sarah Breevoort, M. Herring, Jerome Kurent, I-Hweii Chen, Tuan Vu, Kyle Wuthrich, Nicola Donato, Mulugeta Gebregziabher, Giuseppe Campobello, Brian Dean, and Jonathan Halford. Evaluation

of MPEG-4 AAC audio codec using EEG and EMG signals for use with DICOM® neurophysiology. *Journal of Medical Systems*, 50, 2026. doi: 10.1007/s10916-026-02451-9.

Ary L. Goldberger, Luis A. N. Amaral, Leon Glass, Jeffrey M. Hausdorff, Plamen Ch. Ivanov, Roger G. Mark, Joseph E. Mietus, George B. Moody, Chung-Kang Peng, and H. Eugene Stanley. PhysioBank, PhysioToolkit, and PhysioNet: Components of a new research resource for complex physiologic signals. *Circulation*, 101(23):e215–e220, 2000. doi: 10.1161/01.CIR.101.23.e215.

Shengpeng Ji, Ziyue Jiang, Wen Wang, Yifu Chen, Minghui Fang, Jialong Zuo, Qian Yang, Xize Cheng, Zehan Wang, Ruiqi Li, Ziang Zhang, Xiaoda Yang, Rongjie Huang, Yidi Jiang, Qian Chen, Siqi Zheng, and Zhou Zhao. WavTokenizer: An efficient acoustic discrete codec tokenizer for audio language modeling. *arXiv preprint arXiv:2408.16532*, 2025.

Ard Kastrati, Luca Lanzendörfer, Riccardo Rigoni, John Staib Matilla, and Roger Wattenhofer. Adapting neural audio codecs to EEG. *arXiv preprint arXiv:2511.23142*, 2025.

Rithesh Kumar, Prem Seetharaman, Alejandro Luebs, Ishaan Kumar, and Kundan Kumar. High-fidelity audio compression with improved RVQGAN. *arXiv preprint arXiv:2306.06546*, 2023.

Fabian Mentzer, David Minnen, Eirikur Agustsson, and Michael Tschannen. Finite scalar quantization: VQ-VAE made simple. *arXiv preprint arXiv:2309.15505*, 2023.

Jennifer J. Newson and Tara C. Thiagarajan. EEG frequency bands in psychiatric disorders: A review of resting state studies. *Frontiers in Human Neuroscience*, 12, 2019. doi: 10.3389/fnhum.2018.00521.

Binh Nguyen, Dang Nguyen, Wanli Ma, and Dat Tran. Wavelet transform and adaptive arithmetic coding techniques for EEG lossy compression. In *2017 International Joint Conference on Neural Networks (IJCNN)*, pp. 3153–3160, 2017. doi: 10.1109/IJCNN.2017.7966249.

Ted Painter and Andreas Spanias. Perceptual coding of digital audio. *Proceedings of the IEEE*, 88 (4):451–515, 2000.

Jiapu Pan and Willis J. Tompkins. A real-time QRS detection algorithm. *IEEE Transactions on Biomedical Engineering*, BME-32(3):230–236, 1985. doi: 10.1109/TBME.1985.325532.

Ali Shoeb. CHB-MIT scalp EEG database, 2010.

Hubert Siuzdak. Vocos: Closing the gap between time-domain and Fourier-based neural vocoders for high-quality audio synthesis. *arXiv preprint arXiv:2306.00814*, 2024.

Hubert Siuzdak, Florian Grötschla, and Luca A. Lanzendörfer. SNAC: Multi-scale neural audio codec. *arXiv preprint arXiv:2410.14411*, 2024.

N. Sriraam. Quality-on-demand compression of EEG signals for telemedicine applications using neural network predictors. *International Journal of Telemedicine and Applications*, 2011:860549, 2011. doi: 10.1155/2011/860549.

Michael Tangermann, Klaus-Robert Müller, Ad Aertsen, Niels Birbaumer, Christoph Braun, Clemens Brunner, Robert Leeb, Carsten Mehring, Kai J. Miller, Gernot R. Müller-Putz, Guido Nolte, Gert Pfurtscheller, Hubert Preissl, Gerwin Schalk, Alois Schlögl, Carmen Vidaurre, Stephan Waldert, and Benjamin Blankertz. Review of the BCI Competition IV. *Frontiers in Neuroscience*, 6:55, 2012. doi: 10.3389/fnins.2012.00055.

Punnawish Thuwajit, Phurin Rangpong, Phattarapong Sawangjai, Phairot Autthasan, Rattanaphon Chaisaen, Nannapas Banluesombatkul, Puttaranun Boonchit, Nattasate Tatsaringkansakul, Thapanun Sudhawiyangkul, and Theerawit Wilaiprasitporn. EEGWaveNet: Multiscale CNN-based spatiotemporal feature extraction for EEG seizure detection. *IEEE Transactions on Industrial Informatics*, 18(8):5547–5557, 2022. doi: 10.1109/TII.2021.3133307.

Xiaying Wang, Michael Hersche, Michele Magno, and Luca Benini. MI-BMInet: An efficient convolutional neural network for motor imagery brain–machine interfaces with EEG channel selection. *IEEE Sensors Journal*, 24(6):9084–9096, 2024. doi: 10.1109/JSEN.2024.3358402.

Detai Xin, Xu Tan, Shinnosuke Takamichi, and Hiroshi Saruwatari. BigCodec: Pushing the limits of low-bitrate neural speech codec. *arXiv preprint arXiv:2409.05377*, 2024.

Gaowei Xu, Jun Han, Yao Zou, and Xiaoyang Zeng. A 1.5-d multi-channel EEG compression algorithm based on NLSPIHT. *IEEE Signal Processing Letters*, 22(8):1118–1122, 2015. doi: 10.1109/LSP.2015.2389856.

Neil Zeghidour, Alejandro Luebs, Ahmed Omran, Jan Skoglund, and Marco Tagliasacchi. SoundStream: An end-to-end neural audio codec. *arXiv preprint arXiv:2107.03312*, 2021.

Neil Zeghidour, Alejandro Luebs, Ahmed Omran, Jan Skoglund, and Marco Tagliasacchi. SoundStream: An end-to-end neural audio codec. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 30:495–507, 2022. doi: 10.1109/TASLP.2021.3129994.

Yue Zhao, Yuanjun Xiong, and Philipp Krähenbühl. Image and video tokenization with binary spherical quantization. In *International Conference on Learning Representations (ICLR)*, 2025.

Rui-Chen Zheng, Hui-Peng Du, Xiao-Hang Jiang, Yang Ai, and Zhen-Hua Ling. ERVQ: Enhanced residual vector quantization with intra-and-inter-codebook optimization for neural audio codecs. *arXiv preprint arXiv:2410.12359*, 2025.

Yongxin Zhu, Bocheng Li, Yifei Xin, Zhihua Xia, and Linli Xu. Addressing representation collapse in vector quantized models with one linear layer. *arXiv preprint arXiv:2411.02038*, 2025.

A QRS Complex Detection

To better understand the discrepancy between the low reconstruction error observed on MIT-BIH and the degradation in QRS detection at higher compression ratios, Figure 6 examines the reconstructed ECG waveforms around individual QRS complexes. At CR=16×, both codecs closely follow the original ECG over most of the signal, including the isoelectric regions, while introducing relatively small distortions around the QRS complexes. As compression increases to CR=64×, these sharp peaks become substantially attenuated and broadened, despite the surrounding low-amplitude waveform remaining well reconstructed. Since QRS complexes occupy only a small fraction of the ECG signal, these localized distortions have a limited contribution to global PRD but can substantially affect QRS detection.

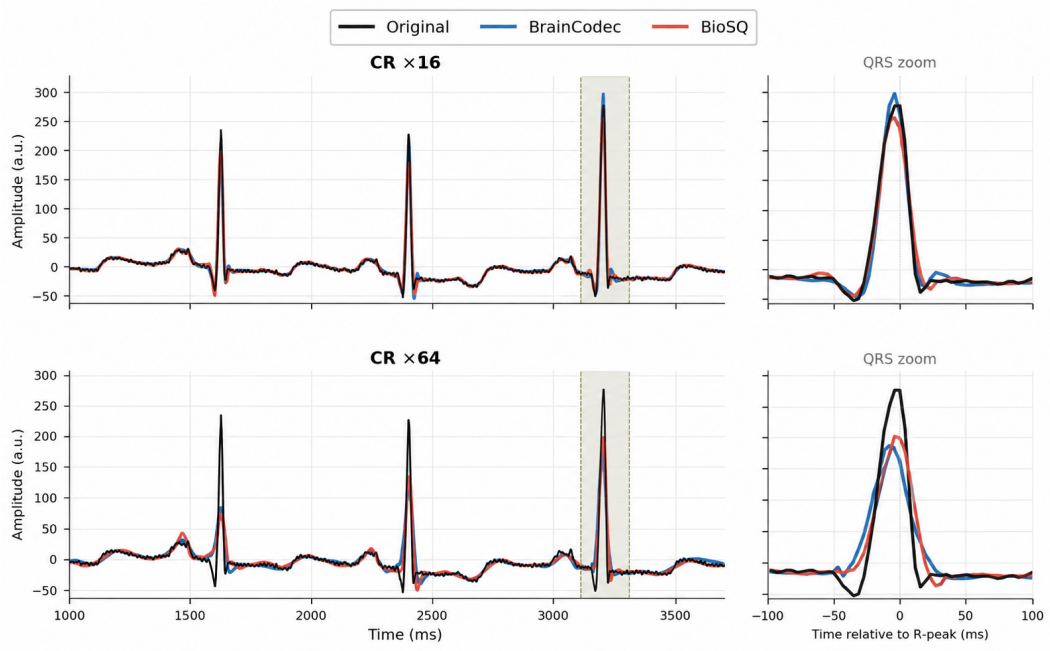
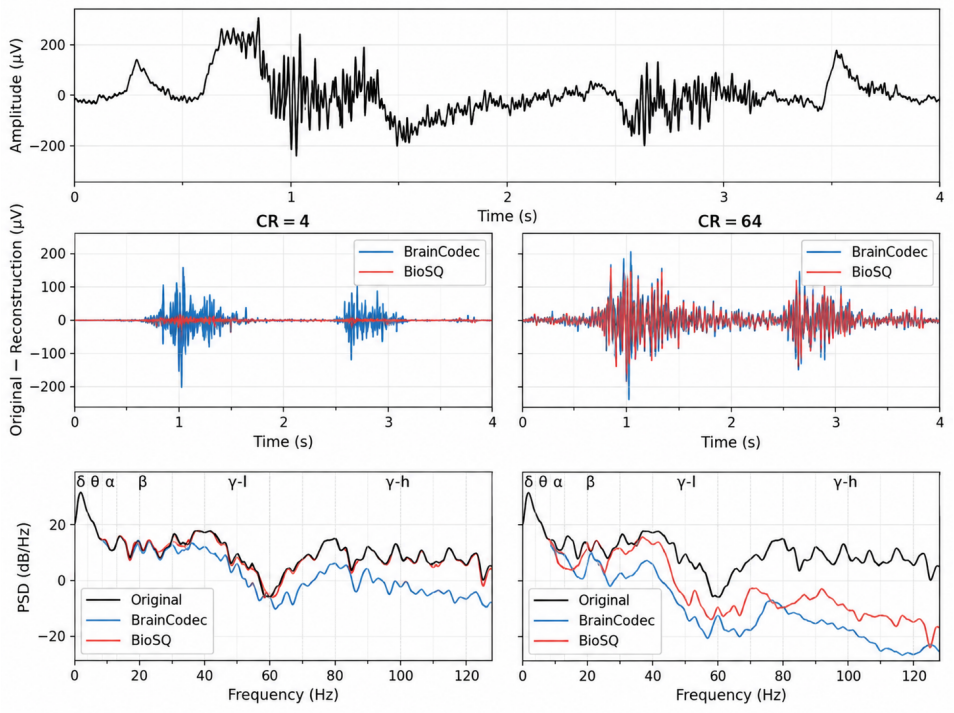


Figure 6: **QRS reconstruction on MIT-BIH at different compression ratios.** Representative ECG segment reconstructed by BrainCodec and BioSQ at CR=16× (top) and CR=64× (bottom). The right panels show an enlarged view around a representative QRS complex highlighted in the corresponding full segment. While both codecs closely reconstruct the low-amplitude regions of the signal, increasing compression progressively distorts the sharp QRS morphology.

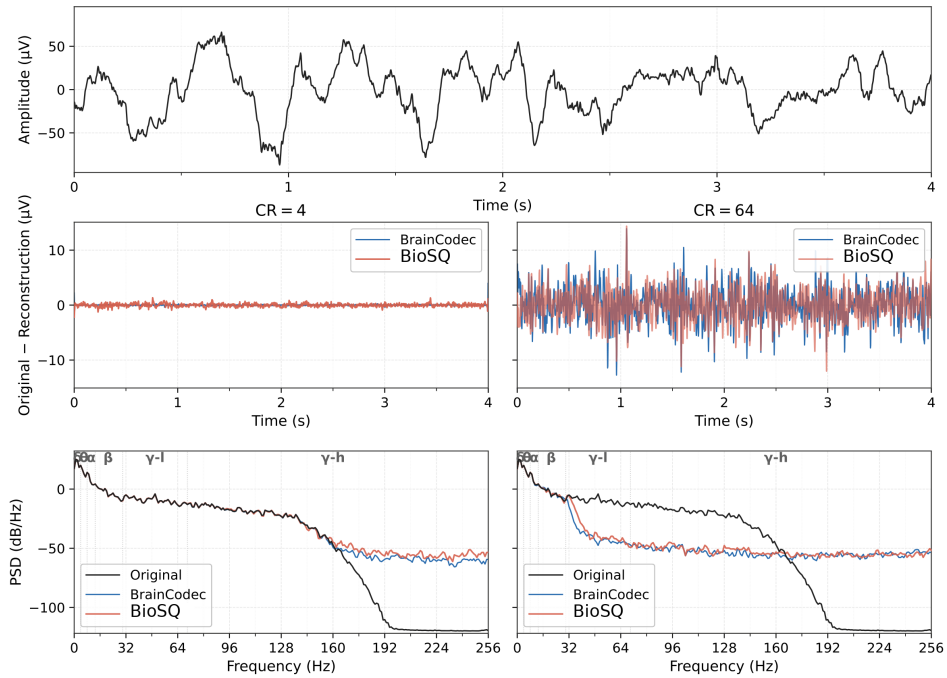
B SWEC and CHB signal reconstruction

Figure 7 shows representative examples of temporal and spectral reconstruction for CHB-MIT (scalp EEG) and SWEC-ETHZ (intracranial EEG) at CR=4× and CR=64×. Across both datasets, increasing the compression ratio produces a clear increase in reconstruction error and a progressive degradation of spectral fidelity. However, the temporal distribution of the error and the frequency range over which spectral information is preserved differ substantially between the two datasets.

For CHB-MIT, at CR=4×, the reconstruction error remains relatively small throughout most of the 4-s window, with the largest deviations concentrated around the high-amplitude transient activity occurring approximately between 1 and 2 s. At CR=64×, the magnitude of the reconstruction error increases substantially and becomes more broadly distributed across the signal. This behavior is also reflected in the spectral domain. At CR=4×, both BrainCodec and BioSQ closely reproduce the low-frequency spectral profile of the original signal, while deviations become progressively larger toward the upper part of the analyzed spectrum (up to approximately 120 Hz). At CR=64×, spectral attenuation is considerably stronger, with both reconstructions underestimating the original spectral power over a much broader frequency range.



(a) CHB-MIT (scalp EEG).



(b) SWEC-ETHZ (intracranial EEG).

Figure 7: Examples of waveform reconstruction and spectral content for CHB-MIT and SWEC-ETHZ. For each dataset, the top row shows the original 4-s signal segment, the middle row shows the reconstruction error for BrainCodec and BioSQ at CR=4× and CR=64×, and the bottom row shows the power spectral density of the original and reconstructed signals at the corresponding compression ratios.

For SWEC-ETHZ, at CR=4×, reconstruction error remains small for both codecs across most of the 4-s window, consistent with the low PRD values observed for this dataset. At CR=64×, the error increases substantially and is distributed throughout the signal rather than being concentrated around transient events, as observed for CHB-MIT. The spectral analysis shows a similar compression-dependent trend: at CR=4×, both codecs closely follow the original spectrum through approximately 150 Hz, with increasing deviations at higher frequencies. At CR=64×, both reconstructions increasingly underestimate the original spectral power between approximately 40 and 160 Hz. Above 160 Hz, however, this trend reverses, with both codecs overestimating the spectral power relative to the original signal. Overall, these results indicate that reconstruction behavior depends on both the compression ratio and the dataset-specific temporal and spectral characteristics of the EEG signals.

C CODEBOOK USAGE AND ARITHMETIC CODING: RVQ VS RBSQ

Figure 8 analyzes how RVQ uses its 256-entry codebooks across EEG (CHB-MIT), iEEG (SWEC-ETHZ), and ECG (MIT-BIH). We report the ranked codeword usage distribution together with the perplexity ratio and the fraction of dead codes, averaged across the four RVQ stages. The perplexity ratio, defined as $2^H/M$, where H is the entropy of the codeword assignment distribution and M is the codebook size, measures how uniformly assignments are distributed across the codebook, whereas the dead-code fraction captures codewords that are never selected. Together, these metrics distinguish between codebooks that remain active but are used unevenly and those in which part of the available discrete capacity is effectively unused.

Codebook utilization varies substantially across modalities and compression ratios. On CHB-MIT, increasing compression progressively concentrates assignments on a smaller subset of codewords: the perplexity ratio decreases from 0.84 at CR=4× to 0.59 at CR=64×, while the dead-code fraction increases from 0.3% to 5.0%. Most codewords nevertheless remain active, indicating that the lower

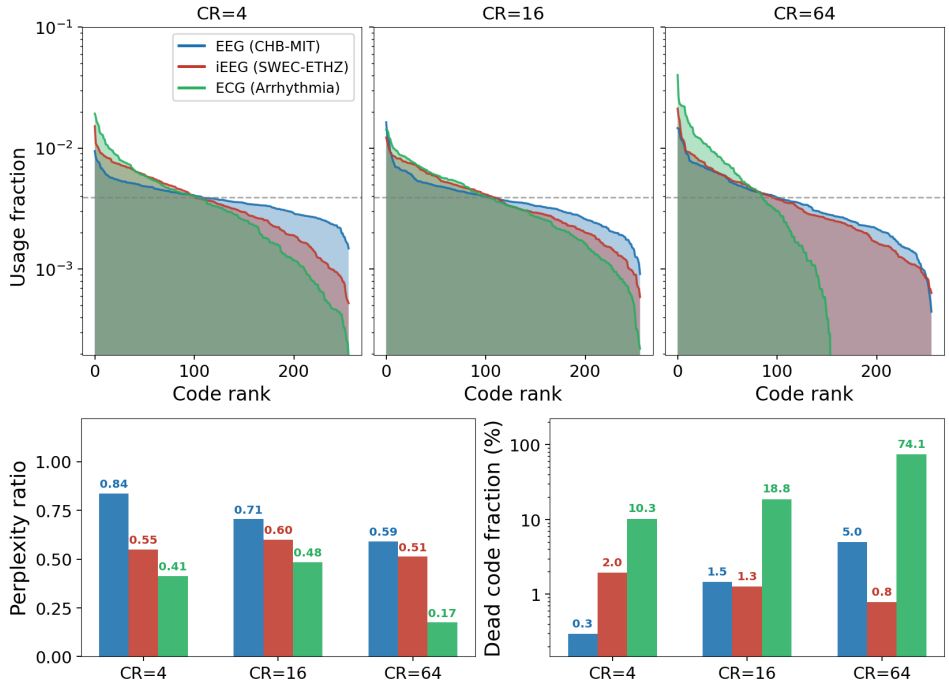


Figure 8: **RVQ codebook utilization across EEG, iEEG and ECG** Top row: usage fraction of the 256 codewords, with the dashed line indicating perfectly uniform usage. Bottom row: perplexity ratio and dead-code fraction, averaged across the four RVQ stages. Higher perplexity indicates a more uniform use of the codebook, whereas the dead-code fraction measures the proportion of codewords that are never selected.

perplexity is mainly driven by increasingly uneven assignment frequencies rather than by widespread codebook collapse. SWEC-ETHZ shows a different behavior, with the perplexity ratio remaining relatively stable across compression ratios (0.55–0.51) and only a small fraction of dead codes. Thus, despite differences in how uniformly codewords are selected, RVQ continues to use most of its available codebook capacity for both EEG and iEEG. A substantially different pattern emerges for ECG, where RVQ exhibits the strongest imbalance in codebook utilization. On MIT-BIH, the perplexity ratio decreases from 0.41 at CR=4× to 0.17 at CR=64×, while the dead-code fraction increases from 10.3% to 74.4%. At high compression, a large fraction of the learned codewords is therefore never selected, and the remaining assignments are concentrated on a relatively small subset of the codebook. Overall, these results show that the effective utilization of learned RVQ codebooks depends strongly on both the recording modality and the compression regime, and that the nominal codebook size can substantially overestimate the discrete capacity that is effectively used.

Figure 9 extends the utilization analysis to RBSQ. Unlike RVQ, RBSQ does not rely on a learned codebook; each 4-bit quantization stage instead produces one of 16 possible binary symbols. We therefore analyze how frequently these symbols are selected and report their ranked usage distribution, perplexity ratio, and fraction of unused symbols. CHB-MIT exhibits the most uniform usage across all compression ratios, with perplexity ratios of 0.89, 0.97, and 0.96 at CR=4×, 16×, and

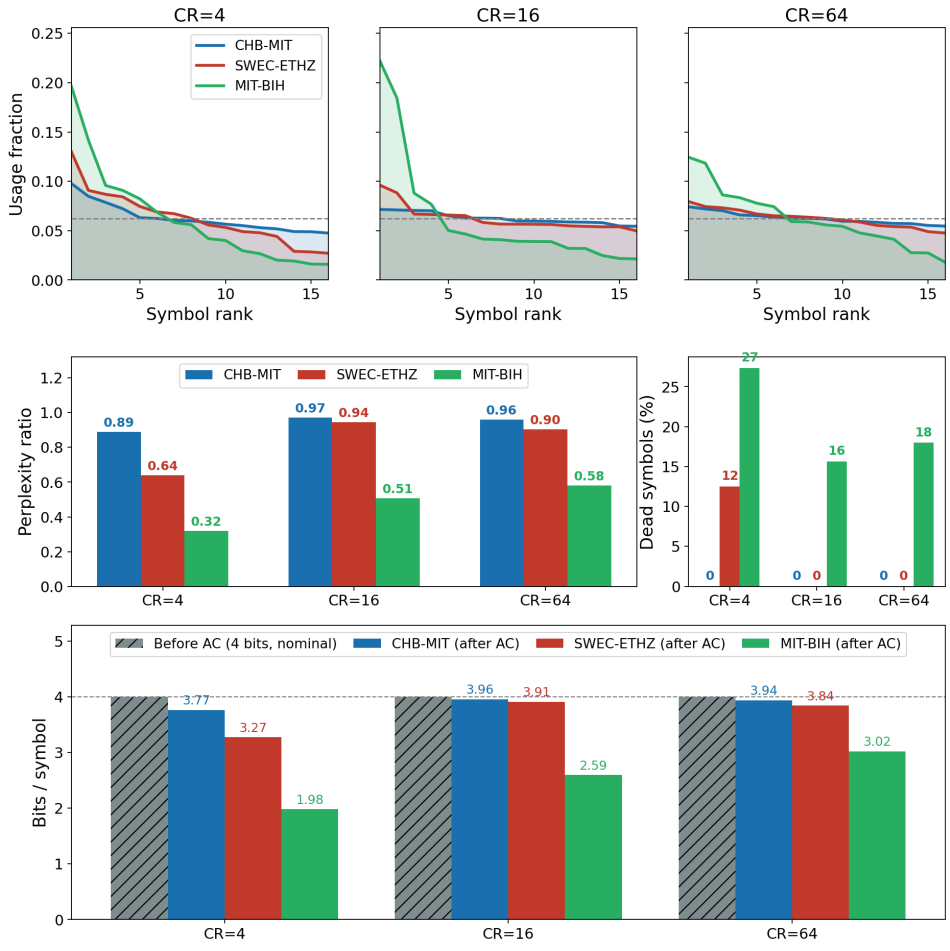


Figure 9: **RBSQ symbol utilization and arithmetic coding gain across CHB-MIT, SWEC-ETHZ and MIT-BIH.** Top row: usage fraction of the 16 possible 4-bit symbols (the same alphabet used by the arithmetic coder), with the dashed line indicating perfectly uniform usage. Middle row: perplexity ratio and dead-symbol fraction, averaged across the eight RBSQ stages. Bottom row: bits per symbol before arithmetic coding (fixed at 4) versus after (the empirical entropy), by dataset and compression ratio.

64×, respectively, and no dead symbols. SWEC-ETHZ is less uniform at CR=4× (perplexity ratio 0.64), where 12% of the symbols are unused, but becomes substantially more balanced at higher compression ratios, reaching perplexity ratios of 0.94 and 0.90 with no dead symbols. MIT-BIH shows the strongest concentration in symbol usage, with perplexity ratios increasing from 0.32 at CR=4× to 0.58 at CR=64×, while 16–27% of the possible symbols remain unused. These differences in symbol utilization are directly reflected in the gains obtained through arithmetic coding. Since each RBSQ symbol requires 4 bits under fixed-length coding, the nearly uniform distributions on CHB-MIT leave little redundancy to exploit, resulting in 3.77–3.96 bits/symbol after arithmetic coding. SWEC-ETHZ achieves a larger reduction at CR=4× (3.27 bits/symbol), but approaches the nominal rate as symbol usage becomes more uniform at CR=16× and 64× (3.91 and 3.84 bits/symbol). The largest gains are observed on MIT-BIH, where the strongly non-uniform symbol distribution reduces the effective rate to 1.98, 2.59, and 3.02 bits/symbol. This explains the substantially larger additional compression obtained through arithmetic coding for ECG compared with EEG and iEEG.

Since BrainCodec is reported at its nominal rate, we also quantify how much arithmetic coding would save on RVQ. Both codecs have the same nominal budget of 32 bits per latent frame (four 8-bit stages for RVQ, eight 4-bit stages for RBSQ), so Figure 10 reports bits per latent frame, computed as the sum of the empirical entropy of all stages. For RBSQ, this is eight times the bits/symbol of Figure 9. On EEG and iEEG the gains are modest for both codecs: RVQ saves 3–12% of the bits and RBSQ 1–18%, with RBSQ saving more only at CR=4×. ECG is the exception. RBSQ saves 51% and 35% at CR=4× and 16×, against 17% and 14% for RVQ, whereas at CR=64× RVQ saves more (32% vs 24%). Arithmetic coding therefore does not systematically favor either quantizer. Since it is lossless, it leaves PRD unchanged, so the comparison at matched nominal rate is not driven by it.

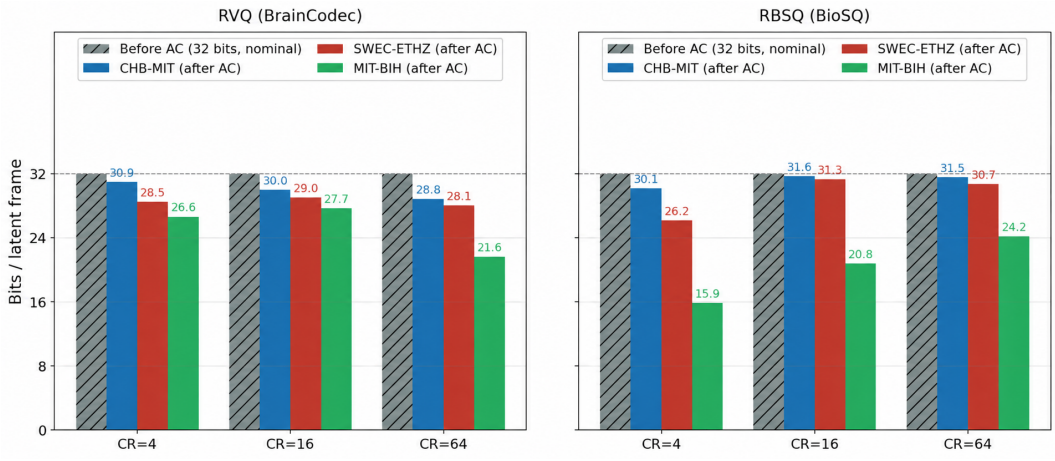


Figure 10: **Arithmetic coding gain for RVQ and RBSQ.** Bits per latent frame before (nominal, 32 bits) and after arithmetic coding (sum of the empirical entropy of all stages), for CHB-MIT, SWEC-ETHZ and MIT-BIH at CR=4×, 16× and 64×.

These results show that arithmetic coding is a useful complement to both quantizers: whenever the learned representation exhibits strongly non-uniform symbol usage, the nominal bit budget overestimates the actual information rate, allowing additional compression without any further reconstruction error.

D Preservation of Physiologically Relevant Signal Features

While PRD provides a general measure of reconstruction fidelity, it does not directly describe whether specific waveform characteristics are preserved. This is relevant for neural signals, where waveform shape can contain physiological information that is not fully captured by spectral measures alone Cole & Voytek (2017). We therefore complement the previous analysis by evaluating the preservation of three morphological properties: amplitude, sharpness and rise-decay asymmetry.

		BrainCodec			BioSQ		
Dataset	Metric	4×	64×	128×	4×	64×	128×
CHB-MIT	Amplitude error (%)	1.00	38.34	45.96	0.66	36.30	43.75
	Sharpness error (%)	1.33	48.85	59.98	0.84	44.56	58.52
	Rise-decay asymmetry	0.01	0.20	0.23	0.00	0.19	0.23
SWEC-ETHZ	Amplitude error (%)	0.39	20.64	25.56	1.16	20.14	24.48
	Sharpness error (%)	1.13	48.95	61.73	3.00	47.28	59.09
	Rise-decay asymmetry	0.01	0.21	0.23	0.01	0.21	0.23

Table 1: **Waveform morphology preservation for BrainCodec and BioSQ on CHB-MIT and SWEC-ETHZ**. Amplitude and sharpness are reported as mean relative errors, while rise-decay asymmetry is reported as the mean absolute difference between the original and reconstructed signals. Metrics are computed over peaks and troughs detected in the original signal. Lower values indicate better preservation of waveform morphology.

Peaks and troughs are first detected in the original signal, retaining only events that stand out from the surrounding signal by at least 5% of the signal standard deviation, thereby reducing the contribution of small fluctuations and noise. The same temporal locations are then evaluated in the reconstructed signal, rather than detecting events independently after reconstruction. This ensures that the comparison measures how compression modifies the morphology of the same waveform events. For each detected event, we evaluate three morphological properties: amplitude, sharpness, and rise-decay asymmetry. Amplitude is measured relative to the median signal level within the corresponding window. Sharpness quantifies how rapidly the signal changes around the peak or trough and is computed from the average magnitude of the slopes on both sides of the event. Rise-decay asymmetry describes differences between the ascending and descending sides of the waveform, with values closer to 0.5 indicating more symmetric slopes. Amplitude and sharpness are reported as relative errors with respect to the original event, whereas asymmetry is reported as the absolute difference between the original and reconstructed signals. Table 1 summarizes the morphological errors obtained for both codecs across compression ratios and datasets.

On CHB-MIT, BioSQ preserves peak and trough amplitudes more accurately than BrainCodec across all compression ratios. At CR=$4\times$, the amplitude error is very small for both methods, reaching 1.00% for BrainCodec and 0.66% for BioSQ. As compression increases, amplitude preservation deteriorates substantially, with errors rising to 38.34% and 36.30% at CR=$64\times$, and to 45.96% and 43.75% at CR=$128\times$, respectively. BioSQ therefore maintains a consistent advantage, although the difference between the two codecs becomes relatively small compared with the overall degradation introduced at high compression. A similar trend is observed for sharpness, with very small errors at CR=$4\times$ that increase markedly as compression becomes stronger. The error rises from 1.33% to 59.98% for BrainCodec and from 0.84% to 58.52% for BioSQ between CR=$4\times$ and CR=$128\times$. BioSQ consistently achieves lower sharpness error, although the difference between the two methods becomes small at the highest compression ratio. At this point, errors close to 60% indicate that the local waveform shape is severely distorted, particularly the slopes that define the sharpness of the detected events. Rise-decay asymmetry shows a different behavior, remaining comparatively well preserved as compression increases. At CR=$4\times$, the error is close to zero for both codecs, and even at CR=$128\times$ it reaches only 0.23 for both methods. Unlike amplitude and sharpness, no clear advantage is observed for either codec, suggesting that the relative balance between the ascending and descending slopes is similarly preserved by both quantization approaches. Figure 11 illustrates these changes for a CHB-MIT event. At CR=$4\times$, both reconstructions closely follow the original waveform, whereas at CR=$64\times$ the differences in amplitude and local shape become clearly visible.

On SWEC-ETHZ, amplitude preservation shows a different behavior from CHB-MIT. At CR=$4\times$, BrainCodec achieves a substantially lower amplitude error than BioSQ, with 0.39% compared with 1.16%. As compression increases, this relationship reverses: at CR=$64\times$, BioSQ obtains a slightly lower error of 20.14% compared with 20.64% for BrainCodec, and the difference increases slightly at CR=$128\times$, with errors of 24.48% and 25.56%, respectively. This crossover follows the same pattern observed in the overall SWEC-ETHZ reconstruction results, where BrainCodec performs better at low compression but BioSQ becomes more effective as the bottleneck becomes more restrictive. Sharpness follows a similar pattern, with BrainCodec achieving lower error at CR=$4\times$ and BioSQ

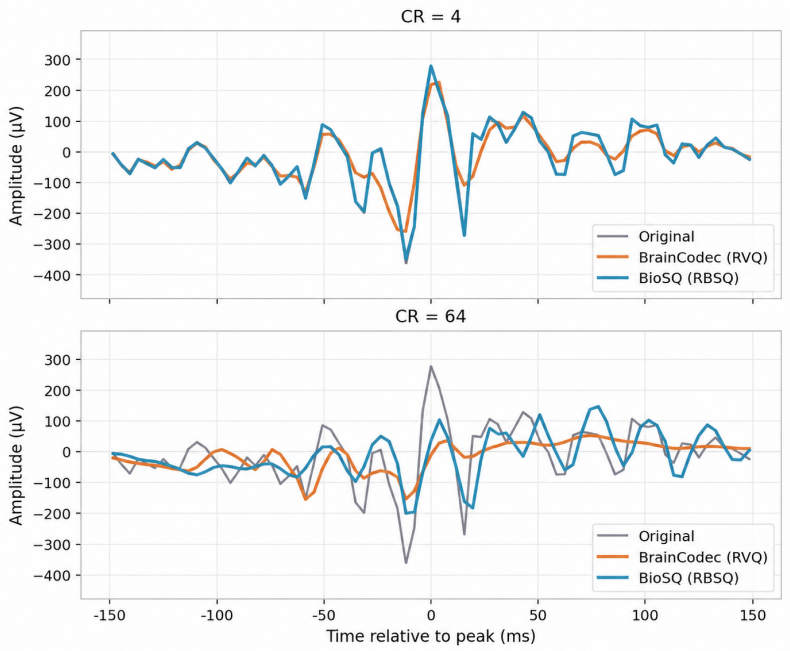


Figure 11: **Example of waveform morphology preservation around a representative event from CHB-MIT**. Original, BrainCodec (RVQ), and BioSQ (RBSQ) signals are aligned at the detected peak ($t = 0$) and shown at CR=4× and CR=64×. At low compression, both reconstructions closely preserve the original waveform, whereas stronger compression produces substantial changes in the amplitude and local shape around the event.

becoming slightly better at the higher compression ratios. The same crossover is therefore observed for both amplitude and sharpness, although the differences between the two codecs remain relatively small once compression increases. Finally, rise-decay asymmetry is again the least affected morphological property, with nearly identical errors for both codecs across compression ratios. This suggests that the relative balance between the ascending and descending slopes is more resistant to compression-induced distortions than amplitude or sharpness.

E Experimental Ceiling

To disentangle the reconstruction error introduced by the encoder–decoder from that caused by quantization, we train the same architecture without a quantization bottleneck, passing the continuous latent representation directly to the decoder. This provides a reconstruction floor for the encoder–decoder at each compression ratio. As shown in Figure 12, the resulting PRD remains below 0.11% at CR=4×, 8×, and 16×, indicating that reconstruction error in these regimes is primarily introduced by RBSQ rather than by the encoder–decoder itself. At CR=64×, however, the reconstruction floor increases to 5.52%, showing that the increasingly restrictive latent representation makes the architecture itself a non-negligible source of error.

We next examine whether the larger reconstruction errors at higher frequencies arise because a full-band model must distribute its limited capacity across frequency bands. To test this, we train separate BioSQ models for each frequency band, giving each model the full 32-bit-per-frame budget while restricting its input to a single band. Figure 12 compares these specialized models with the band-wise PRD of the full-band model at CR=64×.

Specialization substantially reduces reconstruction error across all frequency bands, including β and γ, indicating that limited capacity allocation in the full-band model contributes strongly to the poor reconstruction of high-frequency activity. However, a clear frequency-dependent gap remains: while δ, θ, and α are reconstructed with very low error, PRD remains around 40% in the γ bands even when the entire bit budget is dedicated to them. This suggests that high-frequency content is

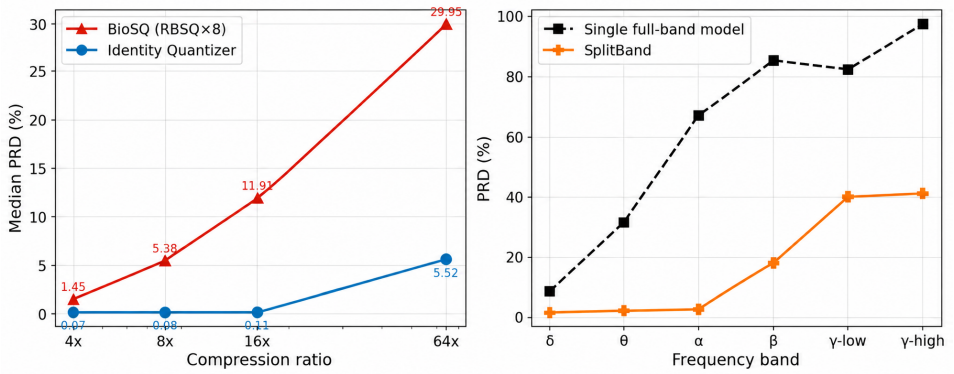


Figure 12: **Experimental ceilings on CHB-MIT** Left: median PRD of BioSQ compared to an Identity Quantizer variant, where the quantization step is bypassed entirely and the encoder-decoder is trained without any constraint to produce a quantizable representation. Right: median PRD per frequency band at CR=64× for a single full-band model versus dedicated specialist models trained and evaluated independently on each band (SplitBand), each with its own full bit budget.

not only disadvantaged by competition for capacity, but also remains intrinsically more difficult to reconstruct under this compression regime.

F RBSQ PARAMETERS ABLATION

We further investigate how the fixed 32-bit-per-frame budget is distributed across the residual stages of RBSQ. Figure 13 compares the 8-stage configuration used in BioSQ with variants using 4 and 16 stages, while preserving the same total bit budget. The number of residual stages has the largest effect at low compression ratios. At CR=4×, the 8-stage configuration achieves substantially lower PRD than both alternatives, whereas the differences narrow at CR=8× and become small at CR=16× and 64×. This suggests that the residual decomposition is particularly important when the latent representation retains sufficient temporal resolution, while under stronger compression the bottleneck imposed by the encoder becomes increasingly dominant.

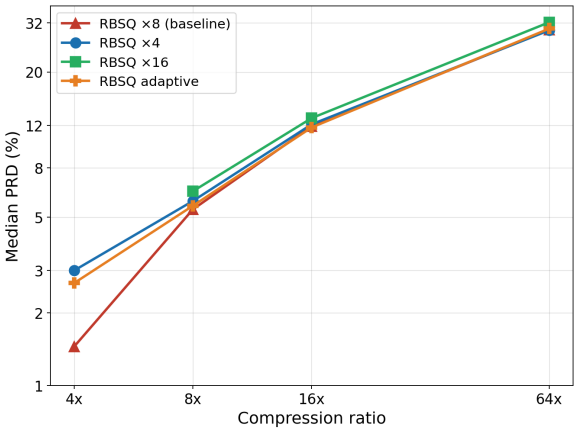


Figure 13: **Ablation of the RBSQ residual configuration on CHB-MIT** Median PRD across compression ratios for different numbers of residual stages and an adaptive bit-allocation strategy, with the total quantization budget fixed at 32 bits per frame. The standard 8-stage configuration performs best at low compression ratios, while differences between configurations become small as compression increases. Adaptive bit allocation does not provide a consistent improvement.

We also consider an adaptive configuration that distributes the 32-bit budget non-uniformly across six residual stages, assigning more bits to stages with larger contributions to reconstruction. Despite this additional flexibility, adaptive allocation does not consistently improve over the uniform 8-stage configuration. It performs similarly at moderate and high compression ratios, but remains substantially worse at CR=4×. These results support the 8-stage configuration used in BioSQ as a robust choice across compression regimes, without requiring compression-dependent or non-uniform bit allocation.

G Extended Architectural Exploration

Figure 14 extends Figure 3 with the two quantizers omitted from the main text, SNAC and RFSQ. The encoder–decoder row is identical to the main figure. The learned-codebook variants behave alike. ERVQ and SNAC remain within about 3 points of RVQ at every compression ratio (46.4% and 49.2% vs. 47.0% at CR=4×), and SNAC is worse at high frequencies at CR=4× (γ_{low}: 75.4% vs. 66.0%). Lookup-free quantizers reduce the error substantially at low compression: at CR=4×, the mean band PRD drops from 47.0% (RVQ) to 21.8% (BSQ), 19.1% (RFSQ) and 11.2% (RBSQ). RFSQ improves on BSQ only in the lower bands at CR=4× and 16×, and is worse overall at CR=16× and 64× (41.2% vs. 37.6% and 69.5% vs. 63.1%). RBSQ has the lowest mean band PRD at all compression ratios. Its advantage narrows at CR=64×, where it is best for δ and θ, whereas BSQ is marginally better for α to γ_{low} and RFSQ for γ_{high}.

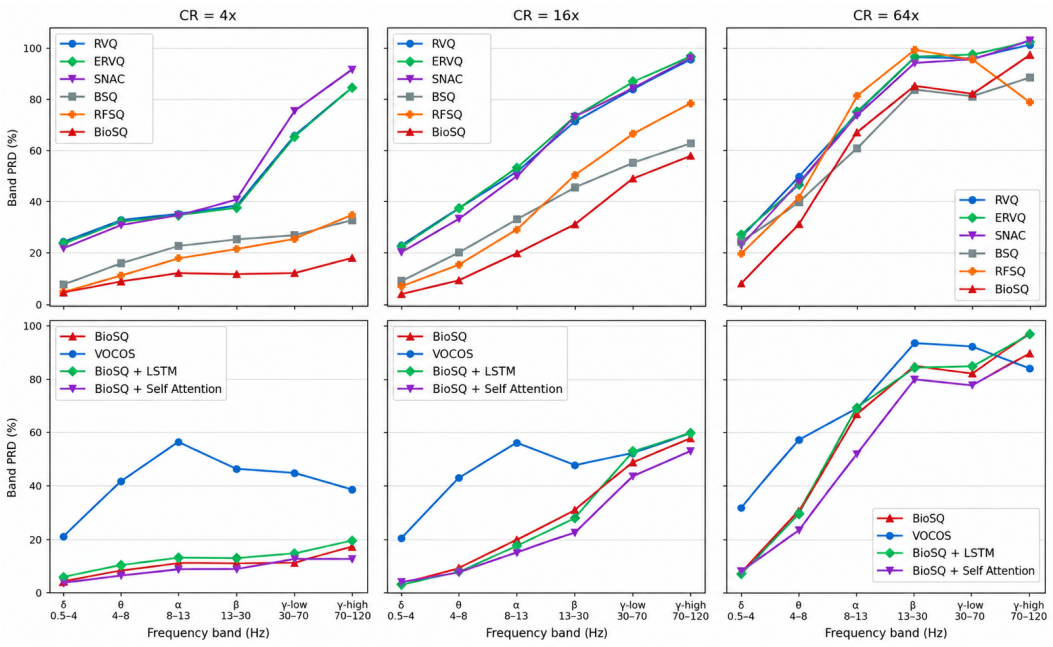


Figure 14: **Band-wise reconstruction error across quantizers and encoder-decoder architectures.** PRD is computed independently over six frequency bands at compression ratios from 4× to 64×. The top row varies the quantizer while keeping the encoder and decoder architecture fixed. The bottom row varies the encoder-decoder architecture while keeping RBSQ, our final quantizer choice, fixed. Lower values indicate better reconstruction fidelity within each frequency band.

H Loss functions

$$\mathcal{L}_{\text{rec}} = \frac{1}{N}\|x - \hat{x}\|_1, \qquad \mathcal{L}_{\text{spec}} = \sum_{i=5}^{10} \frac{\|S_i(x) - S_i(\hat{x})\|_1 + \|S_i(x) - S_i(\hat{x})\|_2}{n_i}, \tag{4}$$

where S_i is the normalized magnitude spectrogram with FFT size 2^i and hop 2^{i-2}, and n_i is its number of elements.