WaveSem: Frequency-Adaptive Tokenization for Disentangling Semantics and Noise in Genomics

Shouzhi Chen[*1,2], **Bing He**[*†1], **Zhenchao Tang**[1], **Jun Zhu**[1]
Minghao Yang[1], **Tianxu Lv**[1], **Yu Wang**[1], **Jiayang Wu**[1]
Fang Wang[1], **Yu Zhao**[1], **Chenchen Qin**[†1], **Jianhua Yao**[†1]
[1]Tencent AI for Life Sciences Lab
[2]Sun Yat-sen University
[*]Equal contribution
[†]Corresponding authors: hebinghb@gmail.com, chenchenqin@tencent.com, jianhua.yao@gmail.com

Abstract

High-throughput scientific sensing, exemplified by genomics, faces a critical representation bottleneck: discrete biological events are often obscured by continuous, high-entropy stochastic noise. Standard neural codecs fail to distinguish deterministic semantics from stochastic sensor noise, inefficiently allocating their information budget to high-entropy fluctuations rather than to low-entropy structural motifs. We introduce WAVESEM, a frequency-adaptive framework that imposes a physically motivated inductive bias on discrete representation learning. Through integrated wavelet decomposition, WAVESEM explicitly disentangles each signal into a structured semantic core and a residual texture component. We validate this approach on large-scale nanopore sequencing data, a domain characterized by extreme sequence lengths and low signal-to-noise ratios (SNRs). WAVESEM achieves a $14\times$ storage reduction while maintaining competitive downstream basecalling accuracy. Moreover, linear probing of the frozen encoder reaches $\mathrm{AUROC} = 0.928$ for 5mC methylation detection, demonstrating that the learned tokens preserve semantic information beyond nucleotide identity. These results establish disentangled vector-quantized representations as a scalable foundation for efficient genomic analysis and long-term archival. Code is available at `https://github.com/GEHAH/WaveSem`.

1 Introduction

The gap between scientific signal acquisition and downstream analysis is widening. High-throughput biological sensing and molecular profiling now generate massive multi-scale datasets, driving rapid advances in representation learning across molecules, single cells, and clinical phenotypes [1, 2, 3, 4], yet our ability to compactly represent raw sequencing waveforms remains limited. A common strategy is to project raw signals directly into task-specific labels, such as basecalls. This end-to-end paradigm is inherently reductive: it discards fine-grained physical cues, including ionic-current amplitudes and dwell times, that are critical for secondary analyses [5]. Conversely, retaining raw waveforms preserves fidelity but incurs prohibitive storage and computational costs. Unlike natural language, genomic sequencing signals contain sparse, low-frequency biological events embedded in high-frequency, high-entropy sensor noise. This structure poses a distinctive representation challenge: how can we learn a compact, high-fidelity discrete representation that preserves the semantic core while remaining efficient for storage and computation [6, 7]?

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

Existing approaches struggle to balance computational tractability with representational fidelity for genomic sequencing signals. Direct raw-waveform analysis is expensive: Convolutional Neural Networks (CNNs) have limited receptive fields for ultra-long dependencies in sequencing data [8], while Transformers scale quadratically with sequence length and are therefore costly for high-frequency inputs [9, 10]. Although structure-adaptive architectures and masked autoencoder representations have proven effective for complex biomolecular modeling [11, 12], neural compression frameworks for continuous waveformssuch as VQ-VAEs [13, 14, 15]typically assume homogeneous information density across the latent space [16, 17, 18]. This assumption conflicts with the physical structure of nanopore signals, where sparse, low-entropy biological events are embedded in dense, high-entropy sensor noise [19, 20]. As a result, generic quantizers can spend limited codebook capacity on stochastic fluctuations, causing codebook under-utilization and eroding fine-grained structures needed for downstream analysis [21].

We propose WAVESEM, a frequency-adaptive framework that imposes a spectral inductive bias on discrete representation learning. Rather than forcing a single codebook to model the full signal bandwidth, WAVESEM combines wavelet decomposition with dual-stream residual vector quantization (RVQ). The input signal is first projected into a latent space and then explicitly decomposed into deterministic semantic components and stochastic texture details via a multi-level discrete wavelet transform (DWT) [22, 23]. This design enables entropy-aware quantization: a Transformer-enhanced semantic stream receives most of the information budget to capture long-range dependencies in the low-frequency band, whereas a lightweight gated texture stream models high-frequency stochastic residuals.

Within a bioinformatics workflow, WAVESEM serves two complementary roles: (1) lossy archival of compact dual-stream RVQ tokens ($\sim$14.2$\times$ smaller than Int16 waveforms under our token accounting), and (2) a reusable discrete representation for downstream models trained on frozen WAVESEM features. Once trained, the tokenizer encodes new nanopore signals out-of-the-box without per-sample optimization, while downstream predictors are trained or calibrated for their target task and deployment domain. WAVESEM is therefore complementary to raw-signal formats and established tools such as Dorado and Remora: we do not position it as a lossless replacement for SLOW5/BLOW5, as a drop-in replacement for raw-signal basecallers, or as evidence that raw waveforms may always be discarded in high-stakes clinical workflows. We validate this disentangled tokenization approach on nanopore sequencing data [7, 24]. Experiments show that WAVESEM achieves a $14\times$ storage reduction, transfers zero-shot across species and GC regimes (Appendix F), and supports competitive token-adapted basecalling as well as epigenetic probing ($\mathrm{AUROC} = 0.928$ on segment-level 5mC, $\mathrm{AUROC} = 0.930$ on per-CpG windows in Appendix G.2, and transfer to endogenous modifications in Appendix G.3). Analysis of the learned latent space reveals a clear topological dichotomy: semantic tokens form distinct, meaningful clusters, whereas texture tokens exhibit broad high-utilization distributions characteristic of high-entropy noise.

Our primary contributions are summarized as follows:

• **Spectral inductive bias for signal tokenization.** We propose WAVESEM, which integrates multi-level wavelet decomposition into the encoder. This spectral inductive bias routes deterministic low-frequency trends and stochastic high-frequency noise through separate processing paths, preventing high-entropy noise from contaminating the semantic latent space.

• **Dual-stream residual vector quantization.** We introduce a frequency-aware tokenization mechanism that discretizes disentangled signal components with differentiated codebook capacity: a deep RVQ preserves the structural integrity of low-frequency semantics, while a shallow RVQ models the high-frequency texture envelope under hierarchical gated fusion.

• **Physically consistent self-supervision.** We formulate a compound objective that incorporates multi-scale spectral constraints and explicit sub-band supervision. We further provide a motivational analysis showing that the sub-band alignment terms protect the semantic approximation while reducing information leakage between streams.

2 Related Work

2.1 Discrete Representation Learning

Neural compression has been substantially advanced by residual vector quantization (RVQ), as demonstrated by audio codecs such as SoundStream [18] and EnCodec [25]. These models progressively refine latent tokens to minimize reconstruction error [13, 14]. However, audio codecs often rely on perceptual masking, discarding details that are imperceptible to humans. This assumption is ill-suited to genomic sequencing signals, where subtle patterns can carry critical physical meaning. Standard VQ frameworks also allocate capacity homogeneously by treating the latent space as having uniform information density [26, 11]. For low-SNR genomic signals, this yields a suboptimal rate-distortion trade-off: limited codebook capacity is spent on high-entropy stochastic noise rather than low-entropy semantic motifs [27, 2]. Our work addresses this limitation by introducing a spectral inductive bias that enforces heterogeneous capacity allocation.

2.2 Deep Learning-Based Feature Extraction from Signals

Deep learning approaches for signal modeling generally fall into two categories: time-domain and frequency-domain paradigms. Time-domain methods typically utilize CNNs for efficient local feature extraction or Transformers for modeling global dependencies. However, these approaches often rely on learned pooling or strided convolutions, which can introduce aliasing during downsampling and mix semantic trends with stochastic noise in the latent space [28, 29, 9]. Conversely, frequency-domain methods leverage transforms like the Discrete Wavelet Transform (DWT) to explicitly separate signal scales [30, 23]. While recent works have integrated wavelets for tasks like denoising or classification [31, 32], they rarely exploit them for discrete tokenization. We bridge this gap by embedding explicit wavelet transforms directly into the quantization bottleneck, enforcing a spectral inductive bias that disentangles the signal spectrum before discretization.

2.3 Genomic Sequencing Signal Compression

High-resolution scientific sensing, typified by nanopore sequencing, presents a critical representation challenge: petabytes of continuous, low-SNR ionic-current streams. Existing compression solutions span a wide spectrum. Lossless approaches include generic codecs such as GZIP and Zstandard, as well as domain-specific tools including VBZ—ONT's official codec combining zig-zag delta encoding with Zstandard—and the SLOW5/BLOW5 SVB-zd format [33, 34]. However, all lossless methods are fundamentally bounded by the Shannon entropy of sensor noise, achieving compression ratios of at most $\sim 2.4\times$. Recently, Jayasooriya et al. [35] proposed ex-zd, a domain-specific lossy method that applies bit truncation and zig-zag delta encoding, reaching $\sim 4\times$ compression by discarding acquisition quantization noise while preserving basecalling accuracy. At the other end of the spectrum, production raw-waveform analysis relies on state-of-the-art ONT callers such as Dorado (`hac/sup`) and research decoders such as Bonito [36, 37, 38] for basecalling, alongside specialized modification callers including Remora, Megalodon, Nanopolish/f5c, and DeepMod2 [39, 7]. While these raw-signal tools establish the field ceiling for nucleotide and site-level modification accuracy, they require storing or streaming bulky raw waveforms, and existing signal compressors treat the waveform as an opaque byte stream rather than learning a reusable representation. We target this complementary gap, using nanopore data as a testbed to show how frequency-adaptive tokenization achieves $\sim 14\times$ compression while exposing discrete features for lightweight downstream probing and token-adapted decoding.

3 Method

Nanopore sequencing signals intertwine two distinct information sources: deterministic base-level semantics and stochastic sensor noise. Uniform quantization treats both identically, wasting bits on noise while under-representing semantics. WAVESEM addresses this mismatch by imposing a spectral inductive bias: a wavelet decomposition splits the latent space into a low-frequency semantic core and multi-scale high-frequency texture residuals, each quantized by a dedicated stream with its own codebook depth and capacity. The overall framework is shown in Figure 1.

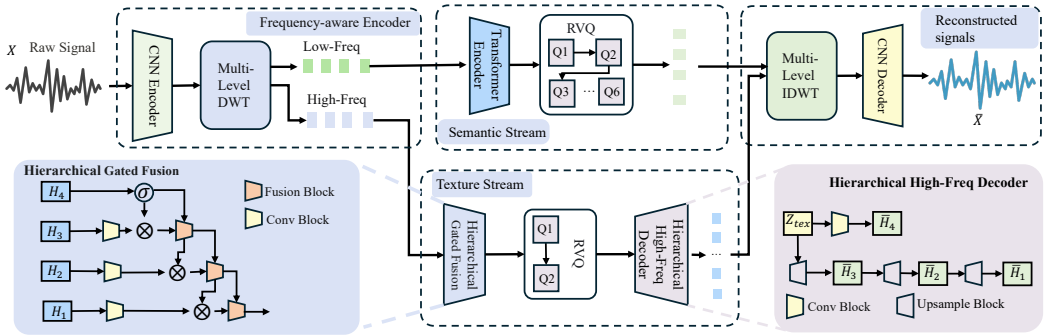


Figure 1: Overview of the WAVESEM architecture. A CNN encoder first extracts latent features, which are then decomposed by a multi-level DWT into a low-frequency semantic core and high-frequency texture components. The **Semantic Stream** contextualizes the core with a Transformer, quantizes it with a 6-layer RVQ, and maps the contextual tokens back to wavelet-coefficient coordinates through a semantic inverse adapter. The **Texture Stream** compresses stochastic details with a 2-layer RVQ guided by a Hierarchical Gated Fusion module, and its decoder is conditioned on the reconstructed semantic core. Reconstruction then applies IDWT and a CNN decoder to produce $\hat{\mathbf{x}}$.

3.1 Problem Formulation

Let $\mathbf{x} \in \mathbb{R}^T$ denote a continuous signal of length T. Our goal is to learn a neural tokenizer that maps $\mathbf{x}$ to a sequence of discrete latent tokens $\mathbf{Z} \in \mathbb{K}^{N \times M}$, where $N \ll T$ is the reduced temporal resolution, M is the number of quantization levels per step, and $\mathbb{K}$ denotes the finite codebook index set.

The framework consists of an encoder $\mathcal{E}$, a discrete quantizer $\mathcal{Q}$, and a decoder $\mathcal{D}$. Given an input $\mathbf{x}$, the reconstructed signal is:

$$\hat{\mathbf{x}} = \mathcal{D}(\mathcal{Q}(\mathcal{E}(\mathbf{x}))) \,. \tag{1}$$

Unlike classical coding schemes that minimize distortion under a fixed rate constraint, our objective balances reconstruction fidelity $d(\mathbf{x}, \hat{\mathbf{x}})$ with representational compactness. The dimensions N and M define the information budget, controlling the temporal resolution and codebook depth of the discrete representation.

3.2 Frequency-Disentangled Encoder

Standard convolutional encoders rely on learned filters to capture multi-scale features implicitly. However, they often struggle to preserve long-range semantic trends while retaining fine-grained local transients, leading to spectral entanglement in the latent space. To address this limitation, we embed a wavelet inductive bias directly into the encoder, disentangling the latent space into coarse semantic structures and fine-scale details before quantization.

Initial Feature Projection. The raw signal $\mathbf{x}$ is first projected into a high-dimensional feature space by a residual convolutional encoder $\mathcal{E}_{\text{cnn}}$ with channel expansion. This stage applies an initial temporal downsampling factor of $S_{\text{cnn}} = n$ (with $n = 4$) to reduce local redundancy and extract robust local features:

$$\mathbf{H}_{\text{pre}} = \mathcal{E}_{\text{cnn}}(\mathbf{x}), \quad \mathbf{H}_{\text{pre}} \in \mathbb{R}^{\frac{T}{n} \times C}. \tag{2}$$

This intermediate representation $\mathbf{H}_{\text{pre}}$ acts as a continuous embedding, preparing the signal for spectral factorization.

Semantic Concentration via Wavelet Decomposition. To promote a frequency-structured partial separation between low-frequency signal semantics and high-frequency residuals, we apply a multi-level DWT using the Daubechies-4 (db4) basis to $\mathbf{H}_{\text{pre}}$. Unlike destructive pooling operations, the DWT provides an invertible decomposition that approximately preserves energy under orthonormal filtering and appropriate boundary handling. Specifically, a J-level decomposition (with $J = 4$) factorizes the features into:

$$\mathbf{z}_{\text{low}}, \left\{\mathbf{z}_{\text{high}}^{(j)}\right\}_{j=1}^{J} = \text{DWT}_J(\mathbf{H}_{\text{pre}})\,. \tag{3}$$

Here, $\mathbf{z}_{\text{low}} \in \mathbb{R}^{\frac{T}{n\cdot 2^J}\times C} = \mathbb{R}^{\frac{T}{64}\times C}$ represents the highly condensed semantic core of the signal, capturing the deterministic low-frequency trends essential for analysis. Conversely, the set $\left\{\mathbf{z}_{\text{high}}^{(j)}\right\}$ encapsulates the detail coefficients at varying scales, primarily encoding texture and noise. This hierarchical structure effectively concentrates meaningful information into a compact latent manifold while maintaining a structured pathway for high-frequency details.

3.3 Dual-Stream Quantization Architecture

3.3.1 Semantic Stream: Contextual Modeling with High-Fidelity Tokens

The low-frequency latent $\mathbf{z}_{\text{low}}$ carries the slowly varying deterministic backbone of the signal. To capture global dependencies within this semantic core, we refine the representation with a Transformer encoder:

$$\tilde{\mathbf{z}}_{\text{sem}} = \text{Transformer}(\mathbf{z}_{\text{low}} + \mathbf{P})\,, \tag{4}$$

where $\mathbf{P}$ denotes Convolutional Positional Encodings (CPE) [40]. This mechanism imparts translation equivariance, enabling the model to generalize across variable-length sequences without rigid positional biases.

Given the high information density of this stream, we employ an RVQ with $N_{\text{sem}} = 6$ stages, allowing the model to progressively refine the semantic tokens:

$$\mathbf{q}_{\text{sem}} = \sum_{m=1}^{N_{\text{sem}}} \mathcal{Q}_m(\mathbf{r}_{m-1})\,, \tag{5}$$

where $\mathbf{r}_0 = \tilde{\mathbf{z}}_{\text{sem}}$ and $\mathbf{r}_m$ is the residual error at stage m. The deep quantization hierarchy enables fine-grained encoding of semantic signal transitions within a compact discrete representation.

3.3.2 Texture Stream: Selective Noise Abstraction

The texture stream processes the multi-scale high-frequency components $\{\mathbf{z}_{\text{high}}^{(j)}\}_{j=1}^{4}$. Because noise exhibits different statistical behaviors across scales, we introduce a Hierarchical Gated Fusion mechanism. Instead of treating these bands independently, we use the coarsest high-frequency band $\mathbf{z}_{\text{high}}^{(4)}$, which captures the most structured texture envelope, to gate and guide the integration of finer stochastic details.

First, we align all finer-scale features ($\mathbf{z}_{\text{high}}^{(3)}, \mathbf{z}_{\text{high}}^{(2)}, \mathbf{z}_{\text{high}}^{(1)}$) to the temporal resolution of the band $\mathbf{z}_{\text{high}}^{(4)}$ using convolutions. The fusion then proceeds in a top-down recursive manner. Let $\mathbf{f}^{(j)}$ denote the fused context at level j. The recursion initializes with $\mathbf{f}^{(4)} = \mathbf{z}_{\text{high}}^{(4)}$. For $j = 3$ down to 1, the feature aggregation is defined as:

$$\mathbf{f}^{(j)} = \Phi_j\left(\text{Concat}\Big[\mathbf{f}^{(j+1)}, \quad \mathbf{z}_{\text{high}}^{(j)\prime} \odot \sigma\Big(\mathcal{G}_j(\mathbf{f}^{(j+1)})\Big)\Big]\right), \tag{6}$$

where $\mathbf{z}_{\text{high}}^{(j)\prime}$ represents the temporally aligned feature at level j, $\mathcal{G}_j$ is a lightweight gating convolution, and Φ_j is a fusion block. The gating mechanism enforces cross-scale consistency: fine-grained noise from $\mathbf{z}_{\text{high}}^{(1)}$ enters the latent code only when it is consistent with coarser-band patterns, thereby suppressing spurious high-frequency artifacts.

The fused output $\mathbf{f}^{(1)}$ is projected and discretized by an RVQ. By anchoring fusion on the coarsest high-frequency component, the model abstracts the noise envelope rather than memorizing independent sample-level fluctuations.

3.4 Reconstruction

Because the discrete tokens are heavily abstracted, the decoder $\mathcal{D}$ must do more than simply invert the encoder: it restores fine-grained details conditioned on the semantic scaffold provided by $\mathbf{q}_{\text{sem}}$.

Texture Synthesis. The quantized texture tokens $\mathbf{q}_{\text{tex}}$ summarize the statistical profile of the high-frequency components. A dedicated hierarchical upsampling network recovers the detail coefficients across scales:

$$\{\hat{\mathbf{z}}_{\text{high}}^{(j)}\}_{j=1}^{J} = \mathcal{D}_{\text{tex}}(\mathbf{q}_{\text{tex}}). \tag{7}$$

Rather than reconstructing each sub-band deterministically, the decoder learns a conditional mapping that restores plausible high-frequency structures consistent with the semantic stream.

Signal Recomposition. Before recomposition, a semantic inverse adapter maps the quantized semantic tokens back to the wavelet-coefficient domain, $\hat{\mathbf{z}}_{\text{low}} = \mathcal{D}_{\text{sem}}(\mathbf{q}_{\text{sem}})$. We then combine this semantic approximation with the synthesized details $\{\hat{\mathbf{z}}_{\text{high}}^{(j)}\}$ via the Inverse Discrete Wavelet Transform (IDWT):

$$\mathbf{H}_{\text{rec}} = \text{IDWT}_J\left(\hat{\mathbf{z}}_{\text{low}}, \{\hat{\mathbf{z}}_{\text{high}}^{(j)}\}_{j=1}^{J}\right). \tag{8}$$

The IDWT enforces correct spectral relationships between bands. A lightweight convolutional post-processor then maps the features back to signal domain:

$$\hat{\mathbf{x}} = \mathcal{D}_{\text{cnn}}(\mathbf{H}_{\text{rec}}). \tag{9}$$

By decoupling the generation of structure and texture while explicitly conditioning texture synthesis on the semantic approximation, this architecture maintains high semantic fidelity for downstream analysis even when the stochastic noise is only approximately reconstructed.

3.5 Learning Objective

To ensure that the learned representation captures both waveform semantics and noise statistics, we optimize the framework end-to-end with the same objective used in our training implementation. The total objective is formulated as:

$$\mathcal{L}_{\text{total}} = \lambda_{\text{time}}\mathcal{L}_{\text{time}} + \lambda_{\text{freq}}\mathcal{L}_{\text{spec}} + \lambda_{\text{sub_low}}\mathcal{L}_{\text{sub_low}} + \lambda_{\text{sub_high}}\mathcal{L}_{\text{sub_high}} + \lambda_{\text{vq}}\mathcal{L}_{\text{vq}}. \tag{10}$$

Reconstruction Fidelity ($\mathcal{L}_{\text{time}}$). While standard ℓ_1 loss encourages pointwise alignment, relying on it alone often leads to over-smoothed predictions. We therefore use it primarily to anchor the signal phase:

$$\mathcal{L}_{\text{time}} = \|\mathbf{x} - \hat{\mathbf{x}}\|_1. \tag{11}$$

Spectral Consistency ($\mathcal{L}_{\text{spec}}$). To preserve energy distributions across scales, we employ the same multi-resolution spectral loss as in training. For a set of STFT configurations $k \in \{1, \ldots, K\}$, let $\mathbf{S}_k = |\text{STFT}_k(\mathbf{x})|$ and $\hat{\mathbf{S}}_k = |\text{STFT}_k(\hat{\mathbf{x}})|$, with $\epsilon > 0$ denoting a small numerical stabilizer. We define:

$$\mathcal{L}_{\text{spec}} = \frac{1}{K}\sum_{k=1}^{K}\left(\frac{\|\mathbf{S}_k - \hat{\mathbf{S}}_k\|_F}{\|\mathbf{S}_k\|_F + \epsilon} + \|\log(\mathbf{S}_k + \epsilon) - \log(\hat{\mathbf{S}}_k + \epsilon)\|_1\right). \tag{12}$$

The first term is a spectral-convergence penalty on the overall magnitude distribution, while the logarithmic term emphasizes low-energy spectral details.

Sub-band Alignment ($\mathcal{L}_{\text{sub_low}}$, $\mathcal{L}_{\text{sub_high}}$). A key risk in dual-stream architectures is information leakage, where the texture stream might attempt to encode semantics (or vice versa). We therefore supervise the low-frequency approximation and high-frequency detail coefficients separately:

$$\mathcal{L}_{\text{sub_low}} = \|\mathbf{z}_{\text{low}} - \hat{\mathbf{z}}_{\text{low}}\|_1, \mathcal{L}_{\text{sub_high}} = \frac{1}{J}\sum_{j=1}^{J}\omega_j\|\mathbf{z}_{\text{high}}^{(j)} - \hat{\mathbf{z}}_{\text{high}}^{(j)}\|_1, \tag{13}$$

where the high-band weights are set to $\omega = [0.5, 1.0, 2.0, 4.0]$ from fine to coarse detail bands. Since $\lambda_{\text{sub_low}} > \lambda_{\text{sub_high}}$, the objective protects the semantic approximation more strongly than stochastic detail bands while still discouraging cross-stream leakage. A detailed analysis is provided in Appendix A.

Discrete Codebook Learning ($\mathcal{L}_{\text{vq}}$). The standard VQ objectivecomprising codebook loss and commitment lossis applied independently to both the semantic and texture RVQ, so that each stream develops a vocabulary tailored to its signal component.

4 Experiments

All datasets, preprocessing details, and baseline configurations are provided in Appendix D.

4.1 Signal Fidelity and Spectral Integrity

Quantitative Benchmarking. We compare the reconstruction fidelity of WAVESEM against a standard VQ-VAE baseline across temporal downsampling factors on a fixed-seed 90/10 segment-level held-out split of HG002 (Appendix D). Here, the downsampling factor (e.g., $64\times$) denotes the ratio between the raw input length and the length of the learned discrete token sequence. As shown in Figure 2(a), the baseline struggles to maintain fidelity as the latent sequence becomes shorter. In contrast, WAVESEM uses its frequency-decoupled architecture to compress semantic content into a reduced temporal budget, achieving an MSE of 0.0757. Beyond in-distribution segments, zero-shot evaluations in Appendix F (Tables 6–7) show that the frozen tokenizer transfers out-of-the-box to an unseen organism (*E. coli*, $\mathrm{MSE} = 0.0830$, Pearson $r = 0.930$), while cross-chemistry transfer from R10.4.1 to R9.4.1 incurs moderate degradation without collapse.

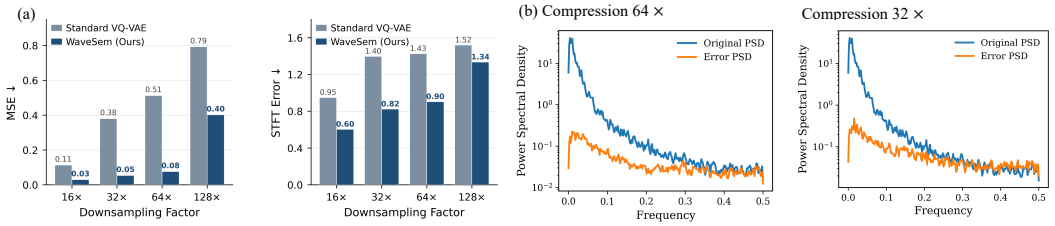


Figure 2: Qualitative and spectral analysis of reconstruction fidelity. (a) Reconstruction fidelity across temporal downsampling factors. (b) Power spectral density (PSD) plots comparing the original signal energy with the reconstruction-error energy.

Qualitative and Spectral Analysis. This structural preservation is further supported by the power spectral density (PSD) analysis in Figure 2(b), which reveals distinct spectral shaping of the quantization error. To visually examine this frequency-structured separation, Figure 6 in Appendix B decomposes the reconstruction into its constituent streams: the semantic stream captures step-like ionic-current plateaus, whereas the texture stream captures structured high-frequency residuals and sensor fluctuations (quantified across 5,000 held-out segments in Appendix C.1, Table 4). Additionally, Figure 5 in Appendix B provides overlaid signal segments and residual curves, showing that the reconstructed morphology closely matches the original signal and preserves step-like transitions.

4.2 Semantic Utility of Discrete Representations

To quantify the information density of the learned representations, we evaluate downstream basecalling performance using a Bonito-style CRF decoder as a controlled reference architecture (whereas production ONT basecalling is represented by Dorado `hac/sup`; see Appendix E.2, Table 5). For a rigorous benchmark among neural codecs, we compare WAVESEM with two baselines: a standard Uniform VQ model and SoundStream. SoundStream was retrained from scratch on the same nanopore dataset as WAVESEM, ensuring that performance differences arise from architecture rather than domain shift.

As shown in Table 1, WAVESEM demonstrates strong semantic retention under aggressive compression. In this setup, the downstream decoder receives post-quantization codebook embeddings summed across RVQ stages and streams, upsamples the $L/64$-length token sequence via a strided transposed-convolutional stack ($\times 8$), and applies a Bonito-style CRF head with beam search

Table 1: Comparison of downstream semantic utility (Basecalling Accuracy) and computational efficiency. WAVESEM tokens retain significantly more biological information than generic or homogeneous baselines at equivalent downsampling levels.

Representation	Input Type	Downsampling	ACC (%) ↑
Bonito	Continuous Waveform	1×	98.12 ± 0.05
SoundStream	Discrete Tokens	64×	92.64 ± 0.41
Standard VQ	Discrete Tokens	32×	73.92 ± 0.54
	Discrete Tokens	64×	65.13 ± 0.91
WAVESEM	Discrete Tokens	32×	**96.88** ± 0.09
	Discrete Tokens	64×	**95.61** ± 0.09

to emit variable-length nucleotide calls without assuming a one-to-one token-to-base alignment (architectural details in Appendix E, Section E.2). At downsampling rates of $32\times$ and $64\times$, our method achieves accuracies of 96.88% and 95.61%, respectively, approaching the raw-waveform Bonito baseline (98.12%), whereas the retrained SoundStream reaches 92.64% at $64\times$. To separate raw waveform reconstruction fidelity from the capacity of the token-adapted CRF decoder, Appendix H (Table 15) also evaluates an independently trained raw-signal Bonito model directly on reconstructed waveforms $\hat{\mathbf{x}}$ (clean 97.19% vs. WAVESEM 85.00%), confirming that while reconstructed waveforms alone incur lossy degradation, pairing discrete tokens with a lightweight token-adapted decoder recovers most basecalling utility.

4.3 Epigenetic Information Probing: 5mC Methylation Detection

Beyond basecalling, we investigate whether the learned representations retain finer-grained epigenetic cues that do not alter the nucleotide sequence itself. We evaluate on the ONT official E. coli 5mC CpG benchmark (Appendix D) using a frozen-encoder probe: the encoder and quantizer are frozen, post-quantization discrete tokens are used as input, and a lightweight MLP head is trained under a shared split, schedule, and seed. Because enzymatic CpG methylation physically shifts ionic current and DM_MSssI/DM come from distinct runs, we also include an uncompressed *Random Frozen CNN* control (frozen randomly initialized encoder) so that scores are interpreted in terms of relative information retention rather than standalone detection accuracy.

Table 2: 5mC CpG methylation detection via linear probing on frozen encoders ($64\times$ downsampling). All models share the same classification head and training protocol; only the frozen encoder differs.

Frozen Encoder	**ACC** ↑	**F1** ↑	**AUROC** ↑	**AUPRC** ↑
Random Frozen CNN (Control)	0.811	0.812	0.889	0.897
SoundStream	0.733	0.743	0.808	0.813
RVQ-VAE	0.812	0.814	0.886	0.901
Single-Codebook VQ-VAE	0.836	0.835	0.914	0.924
WAVESEM (No DWT)	0.779	0.790	0.859	0.870
WAVESEM (Texture Only)	0.791	0.795	0.864	0.877
WAVESEM	**0.852**	**0.851**	**0.928**	**0.936**

Table 2 highlights three findings. First, while the uncompressed random-feature control already reaches 0.889 AUROC due to physical current shifts, WAVESEM's $14\times$-compressed discrete tokens achieve the highest AUROC (0.928, $\Delta = +0.039$ over random continuous features) and outperform Single-Codebook VQ-VAE (0.914), RVQ-VAE (0.886), and SoundStream (0.808). Second, removing the DWT (No-DWT, 0.859) or using SoundStream (0.808) drops performance below the uncompressed random control, showing that generic or un-disentangled quantization discards methylation-relevant variations whereas spectral disentanglement lifts quantized tokens above generic continuous features. Third, we extend this evaluation to per-CpG windows aligned to reference loci and compare site-level methylation estimates with EM-seq and DeepMod2 ($r = 0.949$ and $r = 0.944$; Appendix G.2, Tables 9–10), as well as to endogenous modifications in wild-type *E. coli* (Appendix G.3, Table 11).

4.4 Codebook Utilization Analysis

To validate the effectiveness of disentanglement, we analyze the codebook utilization patterns (entropy profiles) of the learned discrete tokens. The 6-layer Semantic RVQ reveals a coarse-to-fine structure (Figure 3(a,b) and Figure 7 in Appendix C). The first layer (Q_1) exhibits a nearly uniform distribution, indicating that it captures high-entropy biological motifs. Subsequent layers (Q_{2-6}) show progressively steeper, heavy-tailed distributions, confirming a transition toward sparsity. In contrast, the Texture Stream maintains high utilization across all layers. This validates our physical premise: high-frequency components are dominated by stochastic sensor noise, which is inherently non-deterministic and high-entropy.

To investigate how WAVESEM organizes its dual streams in latent space, we visualize the quantized embeddings of both streams using t-SNE (Figure 3(c)) and quantify cross-stream dependence and spectral statistics on 5,000 held-out segments in Appendix C.1 (Table 4). In Figure 3(c), the semantic stream forms a multi-cluster topology reflecting recurring k-mer current motifs, whereas the texture stream follows a broader, more diffuse distribution. Quantitatively, the stored discrete token indices of the two streams are nearly independent (first-layer RVQ normalized mutual information $\mathrm{NMI} = 0.045$, per-timestep feature $\mathrm{CKA} = 0.065$), while the texture reconstruction exhibits a spec-

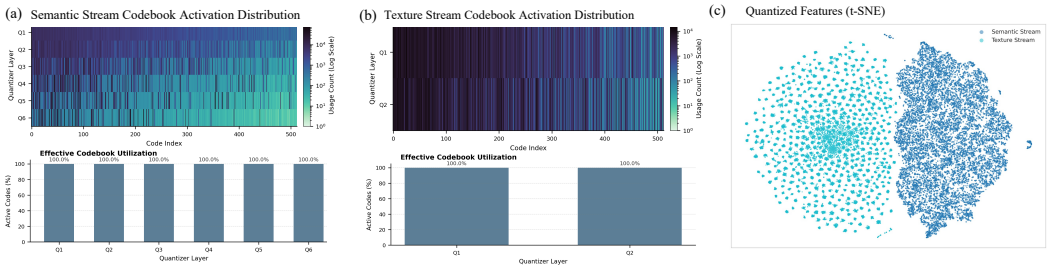


Figure 3: (a,b) Heatmaps of codebook activation counts. Darker colors indicate higher usage. The absence of white vertical stripes indicates that no codebook collapse occurs. (c) t-SNE visualization of the quantized latent space.

tral flatness of 0.028confirming a frequency-structured partial separation into a semantic core and a structured high-frequency residual rather than idealized white noise. This supports our asymmetric quantization strategy, which allocates deeper codebook capacity to the low-frequency semantic stream.

4.5 Ablation Studies

Comparison and analysis of different modules. Table 3 validates both our architectural choices and the decoupled stream design. First, the Single-Stream baseline performs poorly, indicating that a monolithic codebook cannot preserve semantic information under aggressive compression. A basic dual-stream setup without DWT restores accuracy but yields poor reconstruction, showing that explicit spectral bias is required for archival fidelity. Removing Contextual Gating or swapping bit allocations (ACC 92.10%, MSE 0.1845) further confirms the importance of adaptive fusion and asymmetric compression. Second, stream-wise ablations reveal distinct functional roles: the Semantic-Only variant maintains high accuracy (95.61%) but doubles the reconstruction error (MSE 0.1523), whereas the Texture-Only variant collapses across all metrics. These results indicate that the Semantic Stream provides the structural scaffold, while the Texture Stream supplies high-frequency details needed for the full WAVESEM model to achieve strong fidelity and accuracy.

Table 3: Unified component-wise ablation. We evaluate the impact of architectural inductive biases by systematically ablating key mechanisms.

Model Variant	Spectral Bias (DWT)	Dual Stream	Contextual Gating	Bit Split	MSE ↓	STFT ↓	ACC (%) ↑
Single-Stream	×	×	×	×	0.5125	1.4265	65.13
Dual (No DWT)	×	✓	✓	✓	0.2406	1.1602	95.20
Dual (No Gate)	✓	✓	×	✓	0.1250	1.0161	94.80
Dual (Swap Bits)	✓	✓	✓	×	0.1845	1.1032	92.10
WAVESEM (All)	✓	✓	✓	✓	**0.0757**	**0.8697**	**95.80**
WAVESEM (Texture Only)	✓	✓	✓	✓	0.6871	5.8670	81.22
WAVESEM (Semantic Only)	✓	✓	✓	✓	0.1523	1.0847	95.61

Impact of downsampling ratio on semantic stability. To investigate representational robustness, Figure 4(a) plots basecalling accuracy against temporal downsampling factors. WAVESEM exhibits a stable performance plateau: as the downsampling ratio increases from $16\times$ to $64\times$, the accuracy drop remains below 1.3%. In contrast, the Standard VQ baseline suffers a sharp decline at higher compression rates. This confirms that explicitly prioritizing the semantic core helps WAVESEM avoid codebook collapse and retain dense biological information even under a severely constrained representation budget.

4.6 Rate-Distortion and Computational Efficiency Analysis

Information density and storage viability. Figure 4(b) illustrates the information-theoretic efficiency of the learned representation. The standard vector quantization baseline exhibits a rapid drop in fidelity as the compression ratio increases, indicating that uniform tokenization wastes bandwidth on random noise. In contrast, WAVESEM maintains low reconstruction error even at high downsampling levels ($64\times$). This translates into practical storage advantages: on a corpus containing 10 million segments, WAVESEM reduces storage from 190 GB to 13.4 GB, balancing the compactness of lossy codecs with the semantic fidelity required for scientific analysis.

To illustrate the practical significance of WAVESEM for genomics, we provide a multi-axis comparison with general-purpose codecs (GZIP, Zstd, LZMA, BZ2) and domain-specific signal compressors (SLOW5/BLOW5 [33] and ex-zd [35]) in Appendix H (Tables 12–15), covering measured wall-clock encode/decode throughput and peak memory, file-level compression across lossy bit-elimination settings, aggressive-regime rate–distortion against a bit-truncation proxy, and downstream basecalling and 5mC probing under LSB degradation. These experiments clarify that the methods serve distinct operating regimes: lossless SLOW5/ex-zd targets maximal fidelity, lossy ex-zd ($\sim$2.6–4.8$\times$) eliminates least-significant ADC bits with near-intact raw-signal downstream behavior, and WAVESEM ($14.2\times$) targets aggressive storage reduction together with reusable discrete tokens consumed by a token-adapted decoder or lightweight probe.

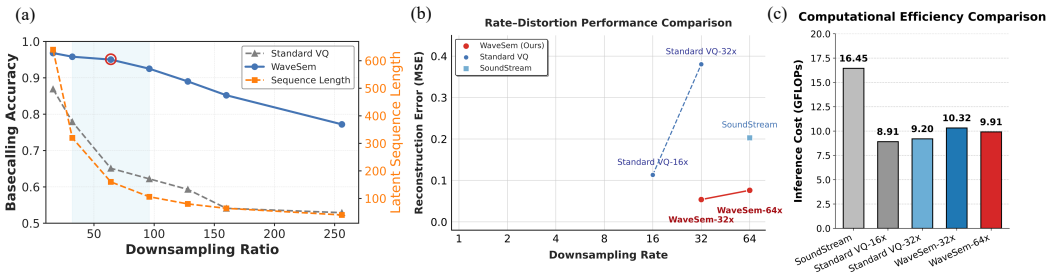


Figure 4: (a) Basecalling accuracy (left axis) and latent sequence length (right axis, orange) as functions of the downsampling ratio. (b) Rate–distortion curve: WAVESEM (red) achieves the lowest reconstruction error at high compression ratios. (c) Computational cost.

Computational throughput from inductive bias. Figure 4(c) compares inference cost in GFLOPs. WAVESEM incurs lower inference cost than SoundStream while achieving a similar storage footprint. Compared with the lightweight Standard VQ baseline, the dual-stream architecture introduces only modest additional overhead. This cost is a favorable trade-off for the substantial gains in reconstruction quality and semantic utility, making WAVESEM a scalable solution for petabyte-scale scientific data.

5 Conclusion

We propose WAVESEM, a frequency-adaptive framework that combines an explicit wavelet inductive bias with dual-stream quantization to learn compact, semantically structured representations of high-resolution genomic sequencing signals. Empirical validation on large-scale nanopore data shows that this physics-informed tokenization approach preserves high-fidelity structural features and competitive downstream utility even at $64\times$ downsampling. These findings suggest that incorporating domain-specific physical constraints into discrete representation learning offers a scalable pathway for managing scientific data at scale, bridging efficient archival and intelligent analysis.

References

[1] Shurjo K. Sen, Eric D. Green, Carolyn M. Hutter, Mark Craven, Trey Ideker, and Valentina Di Francesco. Opportunities for basic, clinical, and bioethics research at the intersection of machine learning and genomics. *Cell Genomics*, 4(1):100466, Jan 2024.

[2] Longtao Zhang, Ruoyu Li, Congrong Ren, Sheng Di, Jinyang Liu, Jiajun Huang, Robert Underwood, Pascal Grosset, Dingwen Tao, Xin Liang, Hanqi Guo, Franck Capello, and Kai Zhao. Lcp: Enhancing scientific data management with lossy compression for particles. *Proceedings of the ACM on Management of Data*, 3(1):1–27, 2024.

[3] Zhenchao Tang, Haohuai He, Shouzhi Chen, Jun Zhu, Tianxu Lv, Jiale Zhou, Jiehui Huang, Yaokun Li, Guanxing Chen, Linlin You, et al. scprototransformer: Scalable reference mapping across molecules, cells, and donors. *Science Advances*, 12(35):eaef0286, 2026.

[4] Zhenchao Tang, Fang Wang, Fan Yang, Jiangning Song, Yiming Li, Jiale Zhou, Yidong Song, Shouzhi Chen, Jun Zhu, Linlin You, et al. sctransmil bridges patient-level disease states and

single-cell transcriptomics for cancer screening and heterogeneity inference. *Nature Communications*, 2026.

[5] Riselda Kodra, Hadjer Benmeziane, Irem Boybat, and William Andrew Simon. Enhancing downstream analysis in genome sequencing: Species classification while basecalling. In *ICLR 2025 Workshop on Machine Learning for Genomics Explorations*, 2025.

[6] Yuk Kei Wan, Christopher Hendra, Ploy N. Pratanwanich, and Jonathan Göke. Beyond sequencing: machine learning algorithms extract biology hidden in nanopore signal data. *Trends in Genetics*, 38(3):246–257, 2022.

[7] Yunhao Wang, Yue Zhao, Audrey Bollas, Yuru Wang, and Kin Fai Au. Nanopore sequencing technology, bioinformatics and applications. *Nature Biotechnology*, 39(11):1348–1365, 2021.

[8] Wei Dai, Chia Dai, Shuhui Qu, Juncheng Li, and Samarjit Das. Very deep convolutional neural networks for raw waveforms. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, page 421425. IEEE, Mar 2017.

[9] Hoai-Chau Tran, Duy Minh Ho Nguyen, Manh-Duy Nguyen, TrungTin Nguyen, Ngan Hoang Le, Pengtao Xie, Daniel Sonntag, James Zou, Binh T. Nguyen, and Mathias Niepert. Accelerating transformers with spectrum-preserving token merging. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.

[10] Liyilei Su, Xumin Zuo, Rui Li, Xin Wang, Heng Zhao, and Bingding Huang. A systematic review for transformer-based long-term series forecasting. *Artificial Intelligence Review*, 58(3), 2023.

[11] Shouzhi Chen, Zhenchao Tang, Linlin You, and Calvin Yu-Chian Chen. Accurate protein–protein interaction prediction: Based on multiview heterogeneous graph autoencoders and random masking. *IEEE Transactions on Neural Networks and Learning Systems*, 2025.

[12] Shouzhi Chen, Zhenchao Tang, Linlin You, Yonghong Tian, and Calvin Yu-Chian Chen. Diconsite: A unified topology-adaptive architecture for protein binding site prediction across ligand modalities. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2026.

[13] Aaron van den Oord, Oriol Vinyals, and koray kavukcuoglu. Neural discrete representation learning. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

[14] Ali Razavi, Aaron van den Oord, and Oriol Vinyals. Generating diverse high-fidelity images with VQ-VAE-2. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d' Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

[15] Long Tung Vuong. Task-driven discrete representation learning. In Yingzhen Li, Stephan Mandt, Shipra Agrawal, and Emtiyaz Khan, editors, *Proceedings of The 28th International Conference on Artificial Intelligence and Statistics*, volume 258 of *Proceedings of Machine Learning Research*, pages 5203–5211. PMLR, 03–05 May 2025.

[16] Marco Pasini, Stefan Lattner, and George Fazekas. Codicodec: Unifying continuous and discrete compressed representations of audio. *arXiv preprint arXiv:2509.09836*, 2025.

[17] Lun Huang, Qiang Qiu, and Guillermo Sapiro. Pq-vae: Learning hierarchical discrete representations with progressive quantization. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern, 2024*, Jan 2024.

[18] 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.

[19] Xinyue Luo, Jin Cheng, Zhaoyi Sun, and Yu Chen. Precision-preserving compression of scientific data: Learn mechanism from data. In *2024 Data Compression Conference (DCC)*, page 572572. IEEE, Mar 2024.

[20] Kleanthis Avramidis, Tiantian Feng, Woojae Jeong, Jihwan Lee, Wenhui Cui, Richard M Leahy, and Shrikanth Narayanan. Neural codecs as biosignal tokenizers. *arXiv preprint arXiv:2510.09095*, 2025.

[21] Hyeongju Kim, Junhyeok Lee, Jacob Morton, Juheon Lee, and Jinhyeok Yang. Improving test-time performance of rvq-based neural codecs. *arXiv preprint arXiv:2509.19186*, 2025.

[22] Ingrid Daubechies. *Ten lectures on wavelets*. SIAM, 1992.

[23] Ashish Kumar, Harshit Tomar, Virender Kumar Mehla, Rama Komaragiri, and Manjeet Kumar. Stationary wavelet transform based ECG signal denoising method. *ISA Transactions*, 114:251–262.

[24] Marc Pagès-Gallego and Jeroen De Ridder. Comprehensive benchmark and architectural analysis of deep learning models for nanopore sequencing basecalling. *Genome Biology*, 24(1):71, 2023.

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

[26] Dianbo Liu, Alex Lamb, Xu Ji, Pascal Junior Tikeng Notsawo, Michael Mozer, Yoshua Bengio, and Kenji Kawaguchi. Adaptive discrete communication bottlenecks with dynamic vector quantization for heterogeneous representational coarseness. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(7):88258833, Jun 2023.

[27] Dan Jacobellis and Neeraja J. Yadwadkar. Learned compression for compressed learning. In *2025 Data Compression Conference (DCC)*, pages 113–122, 2025.

[28] Justin London. Biomedical signal processing: Eeg and ecg classification with discrete wavelet transforms, energy distribution, and convolutional neural networks. *arXiv preprint arXiv:2508.08602*, 2025.

[29] Wangzhuo Yang, Bo Chen, Yijun Shen, and Li Yu. Wavecnns-at: Wavelet-based deep cnns of adaptive threshold for signal recognition. *Applied Intelligence*, 53(23):2881928831, Oct 2023.

[30] Jiangtao Wu, Jiaqi Li, Jie Yang, and Shuli Mei. Wavelet-integrated deep neural networks: A systematic review of applications and synergistic architectures. *Neurocomputing*, 657:131648, Dec 2025.

[31] Weihao Cheng, Yunyun Chen, Bing Tu, Weihao Cheng, Yunyun Chen, and Bing Tu. Phase denoising and unwrapping method for moiré fringes based on multiresolution analysis. *Optics Communications*, 561:130521, 2024.

[32] Muhammad Usman Hadi, Rizwan Qureshi, Ayesha Ahmed, and Nadeem Iftikhar. A lightweight CORONA-NET for COVID-19 detection in x-ray images. *Expert Systems with Applications*, 225:120023, 2023.

[33] Hasindu Gamaarachchi, Hiruna Samarakoon, Sasha P. Jenner, James M. Ferguson, Timothy G. Amos, Jillian M. Hammond, Hassaan Saadat, Martin A. Smith, Sri Parameswaran, and Ira W. Deveson. Fast nanopore sequencing data analysis with SLOW5. *Nature Biotechnology*, 40(7):1026–1029, 2022.

[34] M Kösters, J Leufken, S Schulze, K Sugimoto, J Klein, R P Zahedi, M Hippler, S A Leidel, and C Fufezan. pymzML v2.0: introducing a highly compressed and seekable gzip format. *Bioinformatics*, 34(14):2513–2514, 2018.

[35] Kavindu Jayasooriya, Sasha P Jenner, Pasindu Marasinghe, Udith Senanayake, Hassaan Saadat, David Taubman, Roshan Ragel, Hasindu Gamaarachchi, and Ira W Deveson. A new compression strategy to reduce the size of nanopore sequencing data. *Genome Research*, 35(7):1574–1582, 2025.

[36] Chris Seymour, Oxford Nanopore Technologies Ltd. Bonito: A pytorch basecaller for oxford nanopore reads, 2019. Oxford Nanopore Technologies, Ltd. Public License, v. 1.0.

[37] Shubham Chandak, Kedar Tatwawadi, Srivatsan Sridhar, and Tsachy Weissman. Impact of lossy compression of nanopore raw signal data on basecalling and consensus accuracy. 36(22):5313–5321, 2021.

[38] Qingwen Li, Chen Sun, Daqian Wang, and Jizhong Lou. Basenet: A transformer-based toolkit for nanopore sequencing signal decoding. *Computational and Structural Biotechnology Journal*, 23:34303444, Dec 2024.

[39] Mian Umair Ahsan, Anagha Gouru, Joe Chan, Wanding Zhou, and Kai Wang. A signal processing and deep learning framework for methylation detection using oxford nanopore sequencing. *Nature Communications*, 15(1):1448, 2024.

[40] Xiangxiang Chu, Zhi Tian, Bo Zhang, Xinlong Wang, and Chunhua Shen. Conditional positional encodings for vision transformers. In *International Conference on Learning Representations*, 2021.

[41] Aaron M. Wenger, Paul Peluso, William J. Rowell, Pi-Chuan Chang, Richard J. Hall, Gregory T. Concepcion, Jana Ebler, Arkarachai Fungtammasan, Alexey Kolesnikov, Nathan D. Olson, Armin Töpfer, Michael Alonge, Medhat Mahmoud, Yufeng Qian, Chen-Shan Chin, Adam M. Phillippy, Michael C. Schatz, Gene Myers, Mark A. DePristo, Jue Ruan, Tobias Marschall, Fritz J. Sedlazeck, Justin M. Zook, Heng Li, Sergey Koren, Andrew Carroll, David R. Rank, and Michael W. Hunkapiller. Accurate circular consensus long-read sequencing improves variant detection and assembly of a human genome. *Nature Biotechnology*, 37(10):1155–1162, 2019.

A Motivational Analysis for Spectral Disentanglement

Note: The following analysis provides engineering-level motivations and intuitions for our architectural choices, connecting them to well-established signal processing principles. These are intended as motivational arguments rather than formal proofs.

A.1 Wavelet Decomposition as an Energy-Aware Transform

As described in Section 3.2, the raw signal $\mathbf{x} \in \mathbb{R}^T$ is first mapped to a latent representation $\mathbf{H}_{\text{pre}} \in \mathbb{R}^{(T/n)\times C}$ by a convolutional encoder. A J-level DWT (J=4, Daubechies-4 basis) $\mathcal{W}$ is then applied to $\mathbf{H}_{\text{pre}}$, decomposing it into approximation coefficients $\mathbf{z}_{\text{low}}$ and detail coefficients $\{\mathbf{z}_{\text{high}}^{(j)}\}_{j=1}^{J}$:

$$\mathcal{W}(\mathbf{H}_{\text{pre}}) = \left(\mathbf{z}_{\text{low}}, \mathbf{z}_{\text{high}}^{(1)}, \ldots, \mathbf{z}_{\text{high}}^{(J)}\right). \tag{14}$$

Under an orthonormal DWT implementation with appropriate boundary handling, the db4 transform approximately preserves the latent energy and motivates the following Parseval-style decomposition:

$$\|\mathbf{H}_{\text{pre}}\|_F^2 \approx \|\mathbf{z}_{\text{low}}\|_F^2 + \sum_{j=1}^{J} \|\mathbf{z}_{\text{high}}^{(j)}\|_F^2. \tag{15}$$

This property motivates controlling distortions directly in the wavelet domain, while avoiding reliance on an implementation-independent exact isometry.

A.2 Sub-band Supervision as Sub-band Alignment

Let $\mathbf{H}_{\text{rec}}$ denote the latent feature map reconstructed via IDWT from the reconstructed semantic approximation and decoded texture details (Section 3.4). Under the same orthonormal-transform assumption, the latent reconstruction error admits the following approximate decomposition:

$$\|\mathbf{H}_{\text{pre}} - \mathbf{H}_{\text{rec}}\|_F^2 \approx \underbrace{\|\mathbf{z}_{\text{low}} - \hat{\mathbf{z}}_{\text{low}}\|_F^2}_{\text{Semantic Distortion}} + \sum_{j=1}^{J} \underbrace{\|\mathbf{z}_{\text{high}}^{(j)} - \hat{\mathbf{z}}_{\text{high}}^{(j)}\|_F^2}_{\text{Texture Distortion}}. \tag{16}$$

This decomposition motivates supervising the low- and high-frequency coefficients directly. In the implementation, this supervision is applied through two explicit terms rather than through an additional full-band latent penalty:

$$\mathcal{L}_{\text{sub_low}} = \|\mathbf{z}_{\text{low}} - \hat{\mathbf{z}}_{\text{low}}\|_1, \quad \mathcal{L}_{\text{sub_high}} = \frac{1}{J}\sum_{j=1}^{J} \omega_j \|\mathbf{z}_{\text{high}}^{(j)} - \hat{\mathbf{z}}_{\text{high}}^{(j)}\|_1. \tag{17}$$

Although the ℓ_1 norm differs from the Frobenius-squared form in Eq. 16, it enforces component-wise alignment and is robust to sparse transients. The training weights satisfy $\lambda_{\text{sub_low}}=5 > \lambda_{\text{sub_high}}=1$, so the objective does not incentivize the model to sacrifice semantic fidelity to reconstruct noise. The high-band weights ω_j preserve structurally coherent detail bands while allowing the finest stochastic fluctuations to remain approximate.

A.3 Prevention of Cross-band Leakage

A fundamental risk in dual-stream architectures is cross-band leakage, where high-frequency codebooks inadvertently model low-frequency trends (or vice versa), leading to entangled representations.

Consider the latent reconstruction. Without explicit sub-band supervision, the decoder could minimize the global loss $\|\mathbf{x} - \hat{\mathbf{x}}\|$ by using the texture stream $\hat{\mathbf{z}}_{\text{high}}$ to compensate for errors in the semantic stream, or vice versa. The sub-band losses prevent this by enforcing two complementary constraints. First, $\mathcal{L}_{\text{sub_low}}$ explicitly aligns $\hat{\mathbf{z}}_{\text{low}}$ with the original approximation coefficients after the semantic inverse adapter. Second, $\mathcal{L}_{\text{sub_high}}$ locks each decoded detail sub-band to its corresponding wavelet target. Together, these constraints impose a soft conditional independence structure: the semantic stream must be self-sufficient in reconstructing the approximation coefficients, while the texture stream captures high-frequency residual details conditioned on the semantic scaffold, thereby promoting frequency-structured separation.

A.4 Connection to Rate-Distortion Theory

From an information-theoretic perspective, our framework solves a constrained Rate-Distortion optimization problem. Let $R_{total} = R_{sem} + R_{tex}$ be the total bitrate. Standard codecs optimize distortion D under a constraint $R_{total} \leq C$. Our method introduces a structured distortion measure in the latent wavelet domain:

$$D_{\text{struct}} = D(\mathbf{z}_{\text{low}}, \hat{\mathbf{z}}_{\text{low}}) + \sum_{j=1}^{J} \lambda_j \, D(\mathbf{z}_{\text{high}}^{(j)}, \hat{\mathbf{z}}_{\text{high}}^{(j)}). \tag{18}$$

Although scientific signals in low-SNR regimes may exhibit high energy in high-frequency bands because of noise, semantic information is sparsely concentrated in the low-frequency spectrum. Therefore, maximizing structural fidelity requires allocating bits such that $R_{\text{sem}} \gg R_{\text{tex}}$. The implemented sub-band supervision guides the network toward this operating point on the rate–distortion curve.

B Extended Qualitative Analysis

To further validate the efficacy of our frequency-disentangled representation, we provide additional visual comparisons demonstrating both the internal mechanism of the framework and its reconstruction fidelity across varying degrees of temporal downsampling.

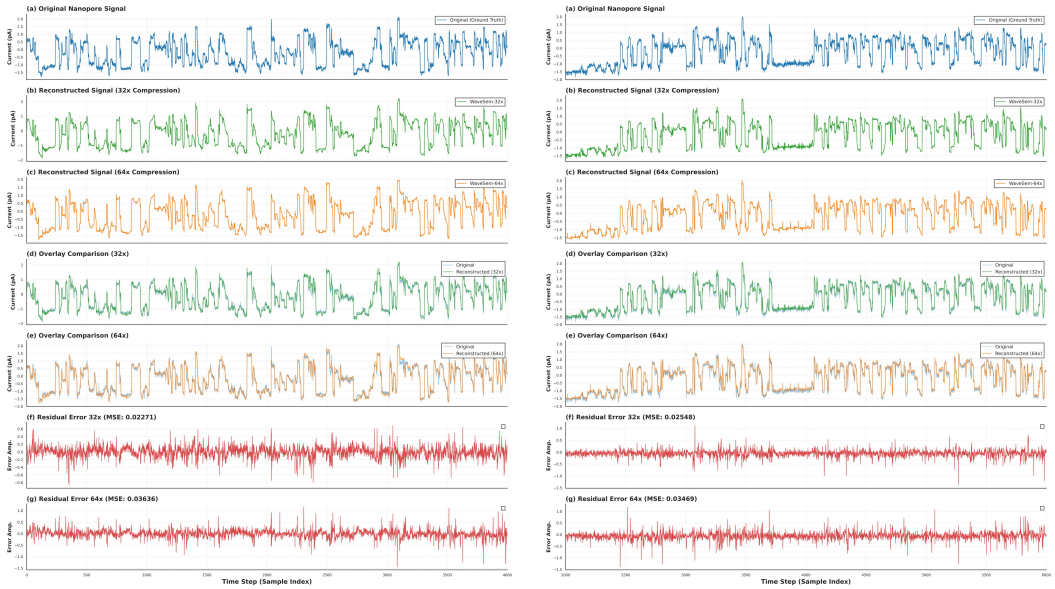


Figure 5: **Detailed reconstruction fidelity analysis across compression levels.** **(a)** Original ground-truth nanopore signal. **(b–c)** Reconstructed waveforms at $32\times$ and $64\times$ downsampling. **(d–e)** Overlay comparison (blue: original; green/orange: reconstructed). The tight overlap shows that phase and amplitude are preserved even when the latent sequence length is reduced by an order of magnitude. **(f–g)** Residual error plots ($\mathbf{x} - \hat{\mathbf{x}}$). The residuals exhibit stochastic characteristics (zero mean and low variance) rather than structured artifacts.

B.1 Visualizing Spectral Disentanglement

WAVESEM is designed to induce a frequency-structured partial separation between low-frequency signal semantics and high-frequency residuals. Figure 6 visualizes this decomposition on two randomly selected nanopore signal segments.

The middle row (Orange) displays the output of the **Semantic Stream**, which captures the step-like ionic-current levels and stable plateaus. The bottom row (Blue) displays the output of the **Texture Stream**, which captures high-frequency transients and structured residual fluctuations rather

than idealized white noise (consistent with its non-zero spectral flatness and residual autocorrelation quantified in Appendix C.1). The top row (Green) shows the final reconstruction combining both streams.

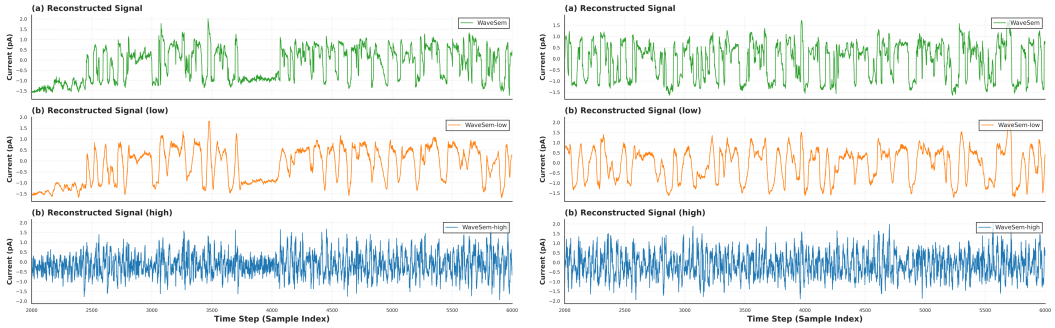


Figure 6: **Visualizing frequency-structured stream decomposition.** We visualize reconstruction components for two genomic signal samples (left and right columns). **(a)** Full reconstructed signal (green), combining both streams. **(b)** *Semantic Component* (orange), reconstructed solely from the low-frequency codebooks. It captures step transitions and structural plateaus essential for downstream analysis. **(c)** *Texture Component* (blue), reconstructed from the high-frequency codebooks. This stream captures high-frequency residual structure and sensor fluctuations. Together, (b) and (c) illustrate a frequency-structured partial separation between the low-frequency semantic scaffold and high-frequency residuals.

C Codebook Utilization and Quantitative Disentanglement Analysis

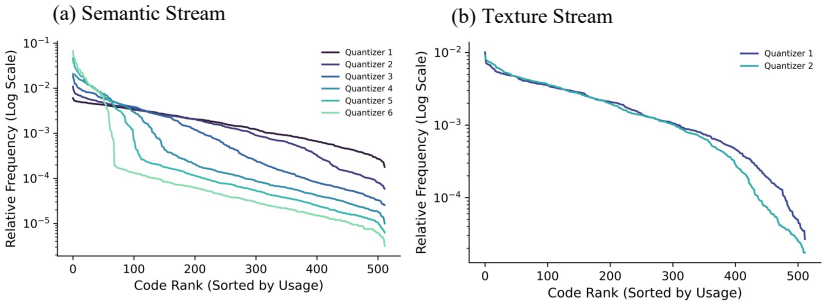


Figure 7: Rank-frequency distributions (log scale). The Semantic Stream (a) shows a progression from high-entropy (flat) to low-entropy (steep), while the Texture Stream (b) shows broad but long-tailed utilization, reflecting stochastic high-frequency structure without codebook collapse.

To examine how the two streams utilize their codebooks, we first analyze the rank-frequency patterns shown in Figure 7:

Semantic Stream: Hierarchical Sparsity. As seen in Figure 7(a), the semantic quantization follows a "coarse-to-fine" trajectory. The first quantizer (Q_1, dark curve) exhibits a near-uniform distribution, indicating high information density required to encode the diverse set of biological motifs. However, subsequent residual layers (Q_{2-6}) transition into heavy-tailed, sparse distributions. This confirms that once the primary semantic structure is captured, the remaining error signal becomes sparse and localized, requiring fewer active codes for refinement.

Texture Stream: Broad Long-tailed Utilization. In contrast, the Texture Stream (Figure 7(b)) maintains broad codebook usage but still exhibits a long-tailed rank-frequency distribution over more than one order of magnitude. This observation is consistent with high-frequency nanopore residuals: stochastic fluctuations require broad capacity, while repeated acquisition and transient patterns create non-uniform usage without claiming a maximum-entropy white-noise codebook.

C.1 Quantitative Disentanglement Metrics

Because qualitative visualizations (Figure 6) and t-SNE projections (Figure 3(c)) on their own are insufficient to establish how strongly the two streams are decoupled, we quantitatively evaluate cross-stream dependence, feature similarity, linear predictability, and residual spectral statistics across $n = 5{,}000$ held-out HG002 segments ($T = 10{,}240$, $200{,}000$ latent positions in total). Table 4 summarizes the results.

Table 4: Quantitative analysis of cross-stream dependence, feature similarity, cross-stream predictability, and spectral statistics on $5{,}000$ held-out segments.

Aspect	**Metric**	**Value**
Cross-stream dependence	RVQ token-index NMI (layer Q_1) $\downarrow$	0.045
	Per-timestep feature linear CKA $\downarrow$	0.065
	Pooled CKA (per-sample normalized)	0.559
	Pooled CKA (raw segment-pooled)	0.970
Cross-stream predictability	Ridge R^2 (per-sample normalized / raw)	≤ 0.68 / ≤ 0.95
Spectral character	Texture stream spectral flatness	0.028
	Residual lag-1 autocorrelation ($\lvert\rho_1\rvert$)	0.573

Table 4 reveals two key findings. First, at the level of the discrete tokens that are actually stored and consumed downstream, the semantic and texture streams are nearly independent: the normalized mutual information between first-layer RVQ token indices is $\mathrm{NMI} = 0.045$ (0.276 nats), and the per-timestep feature CKA across all $200{,}000$ latent positions is 0.065. When features are temporally pooled across an entire $10{,}240$-sample window, raw pooled CKA is high (0.970) and drops to 0.559 after per-sample normalization (with cross-stream Ridge R^2 dropping from ≤ 0.95 to ≤ 0.68), indicating that segment-level coupling is largely driven by shared read-level amplitude and baseline offset rather than token-by-token redundancy. Second, the reconstructed texture stream has a spectral flatness of 0.028 and the overall waveform residual exhibits a lag-1 autocorrelation of 0.573. Thus, the texture stream is not idealized zero-mean white noise; rather, it captures a structured high-frequency residual that retains a fraction of fast ionic-current transients (consistent with the non-trivial Texture-Only 5mC AUROC of 0.864 in Table 2). We therefore characterize WAVESEM's representation as a *frequency-structured partial separation* into a dominant low-frequency semantic core and a structured high-frequency residual.

D Experimental Setup

Dataset, Preprocessing, and Held-out Split Protocol. To validate the scalability and robustness of our framework, we curated a large-scale unlabeled signal corpus for representation learning from the HG002 human genome nanopore sequencing dataset [41]. The continuous ionic current streams were sliced into fixed-length windows of $T = 10,240$ samples, with an overlap of 1,024 points to ensure boundary continuity, resulting in approximately 30 million signal segments. In-distribution reconstruction metrics (e.g., $\mathrm{MSE} = 0.0757$ in Section 4) are computed on a fixed-seed 90/10 held-out split of these HG002 segments; we note explicitly that this partition is made at the *segment level* rather than by read, biological sample, or experimental batch. To evaluate generalization beyond this segment-level split, we additionally test the frozen tokenizer zero-shot on an unseen organism (*E. coli*) and across pore chemistries (R10.4.1 to R9.4.1) in Appendix F. For the downstream basecalling evaluation, we utilized the standard labeled training and validation dataset provided by the Bonito framework [36], which consists of paired raw signal chunks and reference nucleotide sequences.

5mC Methylation Detection Dataset. To evaluate whether the learned representations capture epigenetic information beyond nucleotide identity, we employ the ONT official E. coli 5mC CpG benchmark dataset. This dataset comprises three conditions: (i) DM_MSssI—E. coli K-12 genomic DNA treated with M.SssI methyltransferase, producing fully CpG-methylated signals (positive class); (ii) DM—whole-genome amplified (WGA) DNA lacking CpG methylation (negative

class). We extract a balanced DM_MSssI/DM split for supervised evaluation: 100K training segments (50K per class, split into 85K training and 15K validation) and an independent 20K balanced test set (10K per class). The wild-type (WT) condition, which carries endogenous dam-6mA (GATC) and dcm-5mC (CCWGG) modifications, is not used for CpG-5mC AUROC/AUPRC calculation, and instead serves as the positive class for the native-modification probing experiment in Appendix G.3. All signals are MAD-normalized to zero median and unit scale.

Baselines Configuration. To ensure architectural comparability, the Standard VQ-VAE baseline utilizes the same residual convolutional encoder/decoder backbone as WAVESEM but employs a single, homogeneous codebook ($K = 512$) without wavelet disentanglement. SoundStream [18] was retrained from scratch on the same nanopore corpus. We use a reconstruction-only SoundStream variant by disabling the adversarial discriminator, not because SoundStream is incompatible with 1D signals, but to keep all neural codecs under the same non-adversarial objective used by WAVESEM. This choice may understate the best possible fidelity of a fully tuned SoundStream baseline. We retuned the encoder/decoder strides to match the $64\times$ temporal reduction and selected convolutional receptive fields covering the same input window scale as WAVESEM.

E Implementation Details.

We implemented WAVESEM in PyTorch and conducted training on a high-performance cluster equipped with 8 NVIDIA H20 GPUs. Our training, basecalling, and evaluation code is publicly available at `https://github.com/GEHAH/WaveSem`. The model was optimized using AdamW with a global batch size of 2,048. We employed a cosine annealing learning rate schedule, warming up from 0 to a peak of 1×10^{-4} over the first 5 epochs, followed by a gradual decay. This regime ensures stable convergence of the discrete codebooks.

E.1 Evaluation Metrics Definition

To ensure rigorous benchmarking, we define the metrics used for signal fidelity and semantic utility. Let $\mathbf{x} \in \mathbb{R}^T$ be the ground truth signal and $\hat{\mathbf{x}} \in \mathbb{R}^T$ be the reconstructed signal.

Time-Domain Fidelity. We report the standard Mean Squared Error (MSE), which quantifies the average power of the reconstruction error:

$$\mathrm{MSE}(\mathbf{x}, \hat{\mathbf{x}}) = \frac{1}{T} \sum_{t=1}^{T} (x_t - \hat{x}_t)^2. \tag{19}$$

Lower MSE indicates better preservation of waveform amplitude and phase.

Semantic Utility (Basecalling Accuracy). Basecalling accuracy (ACC) is computed at the nucleotide level by aligning predicted sequences to the reference using standard edit-distance alignment. Let N_{match}, N_{sub}, N_{ins}, and N_{del} denote the number of matched, substituted, inserted, and deleted bases, respectively. Accuracy is defined as

$$\mathrm{ACC} = 1 - \frac{N_{\text{sub}} + N_{\text{ins}} + N_{\text{del}}}{N_{\text{ref}}},$$

where N_{ref} is the length of the reference sequence.

E.2 Model Hyperparameters and Basecalling Decoder Architecture

For all experiments, we set the latent dimension $C = 256$. The codebook size is set to $K = 512$ for all quantizers. The Semantic Stream employs a 4-layer Transformer Encoder with 8 attention heads and a feed-forward dimension of 2048. The DWT uses the fixed db4 basis with $J = 4$ decomposition levels.

Token-to-Sequence Basecalling Architecture. For the downstream basecalling benchmark in Section 4.2, the CRF decoder receives *post-quantization codebook embeddings* (i.e., RVQ-reconstructed continuous vectors in $\mathbb{R}^{N \times C}$) rather than raw integer token indices. Within each

stream, the quantized vectors are summed across RVQ stages ($\mathbf{q}_{\text{sem}} = \sum_{m=1}^{6} \mathcal{Q}_m(\mathbf{r}_{m-1})$ and $\mathbf{q}_{\text{tex}} = \sum_{m=1}^{2} \mathcal{Q}_m^{\text{tex}}(\mathbf{r}_{m-1}^{\text{tex}})$); for the full dual-stream model, the two streams are projected to the shared channel dimension and combined by element-wise addition ($\mathbf{f}_{\text{comb}} = \mathbf{f}_{\text{low}} + \mathbf{f}_{\text{high}}$), whereas semantic-only and texture-only ablations pass a single stream. Given an input window of length L, the fixed-length token sequence of length $N = L/64$ is first upsampled by a strided transposed-convolutional decoder (`decoder_stride = [2, 2, 2]`, providing an $8\times$ temporal expansion to $L/8$). A Bonito-style recurrent/CRF head (5-layer bidirectional LSTM followed by a linear projection) then emits frame-wise state-transition scores trained with the sequence-level conditional random field (CRF/CTC) objective. At inference time, standard Bonito-style CRF beam search decodes these frame-wise transition scores into a variable-length nucleotide string; thus, the output sequence length is not fixed and no one-to-one alignment between discrete tokens and nucleotide bases is assumed.

Production Basecalling Context: Dorado `hac` and `sup`. In Section 4.2, Bonito serves only as a controlled reference decoder architecture to compare discrete representations under identical training conditions, not as a claim to match or replace ONT's production basecaller Dorado. To contextualize our results against current production tools, we benchmarked official ONT Dorado `hac` (`dna_r10.4.1_e8.2_400bps_hac@v4.3.0`, run via Dorado 0.5.3) and the benchmark's super-accuracy (`sup v5.2`) move-table BAMs on 500 primary alignments from the *E. coli* DM R10.4.1 POD5 dataset (Table 5).

Table 5: Production ONT Dorado basecalling accuracy tiers measured on 500 *E. coli* DM R10.4.1 reads.

Dorado Tier	Source	Identity Mean ↑	Median ↑	p10 ↑
`hac@v4.3.0`	Re-run (Dorado 0.5.3)	0.977	0.986	0.953
`sup v5.2`	Benchmark moveBAM	0.989	0.995	0.976
Δ (`sup` − `hac`)	—	+1.2 pp	+0.9 pp	+2.3 pp

Dorado `hac` (0.977 mean identity) and `sup` (0.989 mean identity), together with production modification callers such as Remora, Megalodon, and signal-alignment tools such as Nanopolish/f5c, represent the field ceiling when full raw POD5 waveforms are available. WAVESEM's added value is complementary: it compresses raw signals by $14.2\times$ into discrete tokens that can be archived compactly and queried by lightweight task heads or token-adapted decoders when storing uncompressed waveforms is impractical.

F Zero-Shot Distribution-Shift Reconstruction

Because nanopore ionic current can vary across species, GC composition, and pore chemistries, we evaluate whether a WAVESEM tokenizer trained on human HG002 (R10.4.1) generalizes out-of-the-box.

(A) Cross-Organism Generalization (HG002 → *E. coli*). Both datasets in Table 6 consist of $n = 5{,}000$ native-length segments of $T = 10{,}240$ samples evaluated directly with the frozen $64\times$ HG002 checkpoint under identical preprocessing (`normalize="none"`, matching the pretraining H5 pipeline; seed 42). The relative MSE increase on *E. coli* is modest ($+9.6\%$, from 0.0757 to 0.0830) and waveform correlation remains nearly unchanged (Pearson $r = 0.9343 \rightarrow 0.9302$), indicating that the learned wavelet-domain codebooks capture species-agnostic ionic-current motifs.

Cross-Chemistry and GC-Content Generalization (R10.4.1 → R9.4.1). We further evaluate zero-shot transfer across nanopore pore chemistries under a matched domain- and length-alignment protocol (Table 7). Within R10.4.1, reconstruction quality across low, medium, and high GC-content tertiles is essentially flat (Pearson $r \in [0.930, 0.933]$). Transferring zero-shot from R10.4.1 to the older R9.4.1 chemistrywhich has a different pore constriction geometry and k-mer current profileincurs a larger distortion increase ($\text{MSE} : 0.105 \rightarrow 0.168$, Pearson $r : 0.932 \rightarrow 0.887$) but does not suffer catastrophic collapse. We note that this experiment evaluates waveform reconstruction under distribution shift rather than zero-shot basecalling or methylation calling on R9.4.1.

Table 6: Zero-shot cross-organism reconstruction transfer from human HG002 (in-distribution) to *E. coli* (out-of-distribution) at $64\times$ downsampling.

Condition	**MSE** $\downarrow$	**SNR (dB)** $\uparrow$	**Pearson** r $\uparrow$
HG002 (IID)	0.0757	9.65	0.934
E. coli (OOD, zero-shot)	0.0830	8.98	0.930
Δ (OOD vs. IID)	$+9.6\%$ rel.	-0.67	-0.004

Table 7: Zero-shot cross-chemistry reconstruction transfer from R10.4.1 to R9.4.1 under a matched length/normalization protocol.

Condition / GC Tertile	n	**MSE** $\downarrow$	**SNR (dB)** $\uparrow$	**Pearson** r $\uparrow$
R10.4.1 overall (matched protocol)	2,000	0.105	9.39	0.932
R10.4.1 Low GC $[0.078, 0.375]$	666	0.108	9.34	0.930
R10.4.1 Mid GC $[0.375, 0.448]$	667	0.107	9.47	0.933
R10.4.1 High GC $[0.448, 0.850]$	667	0.098	9.34	0.933
R9.4.1 overall (OOD, zero-shot)	978	0.168	6.78	0.887
R9.4.1 Low GC $[0.078, 0.375]$	212	0.156	6.93	0.891
R9.4.1 Mid GC $[0.375, 0.448]$	260	0.163	6.84	0.888
R9.4.1 High GC $[0.448, 0.850]$	506	0.175	6.69	0.884
Δ (R9.4.1 overall vs. R10.4.1 overall)	—	$+60\%$ rel.	-2.60	-0.045

Within R10.4.1, Pearson correlation is essentially flat across GC tertiles (0.930–0.933); across the R9.4.1 GC bins, Pearson correlation similarly stays within a narrow band (0.884–0.891). Thus, organism and GC-content shifts cause minimal degradation, whereas transferring zero-shot across pore chemistries (R10.4.1 $\rightarrow$ R9.4.1) causes a larger but non-collapsing distortion increase.

G Contextualization with DeepMod2 and Per-CpG Methylation Analysis

To contextualize the epigenetic probing results reported in Section 4.3, we first compare WAVESEM against DeepMod2 [39], a representative deep-learning-based 5mC CpG methylation detection tool, at the segment level, and then evaluate per-CpG site-level resolution and concordance with enzymatic methyl-seq (EM-seq) and DeepMod2.

G.1 Segment-Level Contextualization with DeepMod2

In Table 8, WAVESEM performs per-segment (10,240 samples) classification using a frozen signal-compression encoder that was never trained for methylation detection. Because DeepMod2 produces per-CpG methylation probabilities rather than segment-level labels, we aggregated DeepMod2's per-CpG probabilities across all CpG sites whose aligned signal interval overlaps a given segment and assigned the segment label using the same positive-rate threshold applied to WAVESEM (excluding segments without an aligned CpG site).

Table 8: Segment-level comparison with aggregated DeepMod2 predictions on 5mC CpG methylation detection (†per-CpG DeepMod2 scores aggregated over each 10,240-sample segment).

Method	**ACC**	**F1**	**AUROC**	**AUPRC**
DeepMod2† (segment-aggregated)	0.980	0.981	0.996	0.997
WAVESEM Linear Probe (Ours)	0.852	0.851	0.928	0.936
WAVESEM Fine-tuned (Ours)	0.979	0.979	0.996	0.996

Linear Probing vs. End-to-End Fine-Tuning. When evaluated via a frozen-encoder probe, WAVESEM reaches an AUROC of 0.928 and an ACC of 0.852; when the encoder is unfrozen and

fine-tuned end-to-end on the segment-level task, it reaches ACC = 0.979 and AUROC = 0.996 on this aggregated segment benchmark. Because segment-level classification does not by itself establish single-site resolution or parity with production per-base callers such as DeepMod2 or Dorado-Remora, we next evaluate frozen WAVESEM tokens directly at CpG-centered windows.

G.2 Per-CpG Methylation Evaluation and Concordance with EM-seq and DeepMod2

Alignment and Window-Extraction Protocol. To evaluate retained methylation information at CpG-centered windows without segment averaging, we use Dorado `sup` basecalls with move tables (`mv`/`ts` tags) aligned to the *E. coli* K-12 reference genome (`NC_000913.3`). For each primary alignment with $\mathrm{MAPQ} \geq 20$, we reconstruct the aligned reference sequence from the `MD` tag, locate reference CpG dinucleotides, map the cytosine position to its corresponding raw-signal sample index via the move table, extract a MAD-normalized 10,240-sample window from the POD5 file, and score it with the same frozen 5mC WAVESEM probe used in Section 4.3. Ground-truth construction labels are DM_MSssI = methylated (1) and DM = unmethylated (0). Reference CpG sites are deterministically subsampled (`ref_pos % 150 == 0`) so that both conditions share the same loci (2,010 shared sites, with mean per-site read coverage of $99.4\times$ on DM_MSssI and $91.8\times$ on DM). We emphasize that this is an information-retention probe on CpG-centered windows, not a supervised per-site caller trained on EM-seq and not a competition with production callers such as DeepMod2 or Dorado-Remora.

Table 9: Per-CpG read-level probing (no site aggregation) on CpG-centered 10,240-sample windows (2,010 shared reference CpG loci; $\sim$99$\times$ positive / $\sim$92$\times$ negative read coverage).

Frozen Probe	**Read-level AUROC** $\uparrow$	**AUPRC** $\uparrow$	**ACC@0.5** $\uparrow$	**Pos / Neg Mean** p
Random Frozen CNN (Control)	0.874	0.877	0.797	0.775 / 0.289
WAVESEM	**0.930**	**0.935**	**0.854**	0.821 / 0.250

Table 10: Same-site per-CpG methylation frequency concordance (Pearson r) across shared reference loci against EM-seq ground truth and DeepMod2.

Comparison	**Pearson** r $\uparrow$
Random Frozen CNN $\leftrightarrow$ EM-seq	0.931
WAVESEM $\leftrightarrow$ EM-seq	**0.949**
WAVESEM $\leftrightarrow$ DeepMod2	**0.944**
DeepMod2 $\leftrightarrow$ EM-seq	0.995

As shown in Table 9, the primary unaggregated read-level AUROC on CpG-centered windows is 0.930, matching the segment-level result (0.928) and outperforming the uncompressed *Random Frozen CNN* control by $+0.056$ AUROC. When read-level scores are aggregated at each shared reference CpG and compared against EM-seq and DeepMod2 (Table 10), WAVESEM achieves strong cross-method concordance ($r = 0.949$ with EM-seq vs. 0.931 for `random_frozen`, and $r = 0.944$ with DeepMod2). At the same time, WAVESEM's predicted probabilities are less sharply calibrated than DeepMod2 (positive/negative condition means of $0.821/0.250$ vs. $0.969/0.045$ for DeepMod2), because each 10,240-sample window spans multiple surrounding k-mers rather than isolating a single CpG. We therefore claim methylation information retention and cross-method concordance rather than parity with DeepMod2, Dorado-Remora, or other production methylation callers.

G.3 Native Modification Probing Beyond Enzymatic CpG-5mC

Why retain a compressed signal representation? Basecalls are a model-dependent, lossy summary of the ionic-current trace. Retaining signal-derived information can support appropriately adapted future basecalling or modification models, signal-level quality control, and analyses of modifications not represented in a canonical nucleotide string. WAVESEM targets compact archival and inexpensive re-analysis: it compresses the signal by $\sim$14.2$\times$ into discrete tokens that can be consumed by a lightweight task head without retraining the tokenizer. We do not claim that these lossy

tokens preserve every property of raw POD5 files or that they can replace raw signals in high-stakes clinical settings.

Probing endogenous modifications (WT vs. DM). To test whether the learned discrete tokens transfer beyond enzymatic CpG-5mC methylation, we evaluated a native-modification task on the same ONT *E. coli* benchmark. The positive class is wild-type (WT) *E. coli* K-12 DNA, which carries endogenous dam-mediated 6mA methylation at GATC motifs and dcm-mediated 5mC methylation at CCWGG motifs; the negative class is double-mutant (DM, dam$^-$/dcm$^-$) DNA lacking these endogenous methyltransferases. We froze each codec encoder and quantizer and trained the same lightweight MLP probe used in Section 4.3 (85K training and 15K validation segments, 30 epochs).

Table 11: Native modification probing (WT with endogenous dam-6mA at GATC and dcm-5mC at CCWGG vs. DM dam$^-$/dcm$^-$) using frozen encoders at $64\times$ downsampling.

Frozen Encoder	**ACC** ↑	**F1** ↑	**AUROC** ↑	**AUPRC** ↑
RVQ-VAE	0.526	0.619	0.541	0.525
SoundStream	0.516	0.646	0.543	0.530
Single-Codebook VQ-VAE	0.540	0.643	0.574	0.560
WAVESEM (Semantic)	0.568	0.679	0.662	0.622
WAVESEM-Dual	**0.599**	**0.680**	**0.678**	**0.644**

As shown in Table 11, WAVESEM-Dual improves AUROC by 0.104–0.137 over all non-WAVESEM neural-codec baselines, supporting transfer beyond enzymatic CpG-5mC to endogenous modifications that include 6mA at a distinct sequence motif. Combining the semantic and texture streams (WAVESEM-Dual, AUROC = 0.678) also improves slightly over the semantic stream alone (0.662). Because the WT-versus-DM contrast includes both dam-6mA and dcm-5mC, it does not isolate the contribution of 6mA alone. Moreover, the absolute AUROC (0.678) is modest: native modifications induce subtler current perturbations than dense enzymatic CpG methylation, and read-level WT/DM labels are weak at the segment level because a sampled window need not overlap a modified motif. We therefore interpret these results as evidence of retained native-modification information rather than as a competitive site-level 6mA caller, and we restrict our stated use cases to compact archival, token-adapted re-basecalling, and lightweight probing of modification-related signals.

H Compression Baseline Comparison

To replace literature placeholders with direct measurements on our hardware, Table 12 reports measured compression ratios (relative to Int16, 2 bytes/sample), wall-clock encode and decode throughputs, and execution devices across generic lossless codecs (GZIP, Zstd, LZMA, BZ2), domain-specific nanopore compressors (SLOW5 SVB-zd and ex-zd via official `slow5tools` v1.4.0), and WAVESEM. Generic codecs and WAVESEM are benchmarked on $n = 2{,}000$ held-out H5 segments ($2{,}000 \times 10{,}240$ samples, ∼41 MB Int16 equivalent), while SLOW5 and ex-zd are measured on 300 real *E. coli* DM POD5 reads (∼68 MB of Int16 signal).

Table 12: Measured compression ratios (CR relative to Int16), wall-clock encode/decode throughputs, and execution devices. "Size" estimates are for a corpus of 10M segments (T=10,240, 190.7 GB Int16).

Method	**Type**	**CR (Int16)** ↑	**Size (GB)** ↓	**Encode (MB/s)** ↑	**Decode / RT (MB/s)** ↑	**Device**
Generic Lossless						
GZIP (L6)	Lossless	1.24×	153.8	36.1	177.1	CPU
Zstd (L19)	Lossless	1.29×	147.8	4.1	463.4	CPU
BZ2	Lossless	1.29×	147.8	16.0	25.2	CPU
LZMA	Lossless	1.52×	125.5	3.0	48.3	CPU
Domain-Specific Lossless						
SLOW5 (SVB-zd)	Lossless	2.25×	84.8	24.3	23.2	CPU (`slow5tools`)
ex-zd (Lossless)	Lossless	2.32×	82.2	46.8	24.7	CPU (`slow5tools`)
Lossy						
ex-zd (Lossy, `-b 3`)	Lossy	3.75×	50.9	37.1	27.2	CPU (`slow5tools`)
WAVESEM (Ours)	Lossy neural	**14.2**×	**13.4**	33.2	18.0 (round-trip)	GPU (4.7 GB peak)

Table 12 shows a clear hierarchy: generic lossless codecs remain at $\leq 1.52\times$, domain-specific lossless codecs (SLOW5/ex-zd) reach $\sim$2.3$\times$, lossy ex-zd (`-b 3`) reaches 3.75$\times$, and WAVESEM reaches $\sim$14.2$\times$ while producing directly reusable discrete tokens. On GPU (4.7 GB peak memory), WAVESEM's encoding throughput (33.2 MB/s) is comparable to CPU GZIP (36.1 MB/s) and in the same range as CPU ex-zd (37.1–46.8 MB/s), whereas full waveform round-trip decoding is slower (18.0 MB/s); we therefore report the hardware device per row and do not claim a same-hardware speed advantage.

Because WAVESEM and ex-zd target different operating regimes, we further compare them along three axes: authentic file-level compression across bit-elimination settings, rate–distortion at aggressive compression ratios, and downstream task behavior (including running an existing raw-signal Bonito basecaller on reconstructed waveforms).

(A) Authentic File-Level Compression Across ex-zd Settings. We measured official `slow5tools` v1.4.0 on the 300 *E. coli* DM reads ($\sim$68 MB of Int16 signal) across lossless and lossy bit-elimination settings (`-b 1` to `-b 4`):

Table 13: Measured file-level compression and throughput using official `slow5tools` v1.4.0 on 300 *E. coli* DM reads ($\sim$68 MB of Int16 signal). WAVESEM's normalized-domain MSE is omitted here because it is not directly comparable with integer ADC-domain MSE.

Method	Type	CR (Int16) ↑	Enc. MB/s	Dec. MB/s	ADC-domain MSE
SLOW5 (SVB-zd)	Lossless	2.25×	23.8	19.9	0
ex-zd lossless	Lossless	2.32×	45.9	26.7	0
ex-zd `-b 1`	Lossy	2.64×	32.1	24.7	0.53
ex-zd `-b 2`	Lossy	3.06×	31.2	13.2	1.50
ex-zd `-b 3`	Lossy	3.75×	25.3	26.7	5.33
ex-zd `-b 4`	Lossy	4.81×	47.3	30.6	20.8
WAVESEM	Lossy neural	**14.2**×	—	—	—

These measurements confirm the $\sim$3–5$\times$ compression scale of lossy ex-zd, while WAVESEM achieves a compression ratio approximately 3–4$\times$ higher than these lossy settings.

(B) Rate–Distortion at Aggressive Compression. Official ex-zd combines least-significant-bit (LSB) elimination with a lossless backend and is not designed for WAVESEM's $\sim$14$\times$ regime. To compare waveform distortion in the same normalized floating-point domain on 5,000 held-out segments, we evaluate uniform bit truncation as an explicitly labeled ex-zd-style proxy (rather than a measured ex-zd file size):

Table 14: Rate–distortion comparison on 5,000 held-out normalized segments against a uniform bit-truncation proxy.

Method	Nominal bits/sample ↓	Int16-equiv. CR ↑	Normalized MSE ↓	SNR (dB) ↑
Truncation proxy, 8-bit	8.0	2.0×	0.00002	45.5
Truncation proxy, 6-bit	6.0	2.7×	0.00033	33.4
Truncation proxy, 4-bit	4.0	4.0×	0.00574	20.9
Truncation proxy, 2-bit	2.0	8.0×	0.15288	6.7
WAVESEM	**1.125**	**14.2**×	**0.0757**	—

At the high-compression operating point, WAVESEM uses fewer bits per sample than the 2-bit truncation proxy (1.125 vs. 2.0) while approximately halving normalized MSE (0.0757 vs. 0.1529); interpolating the proxy curve requires $\sim$2.4 bits/sample ($\sim$6.7$\times$ CR) to match WAVESEM's distortion. This indicates a rate–distortion advantage in the aggressive compression regime, whereas in the mild-loss regime ($\leq 4\times$) bit truncation introduces negligible waveform distortion.

(C) Reconstructed-Waveform Basecalling and Downstream Operating Boundaries. To separate waveform reconstruction fidelity from the capacity of a newly trained token→CRF decoder, we also evaluated an independently trained raw-signal CRF Bonito modelwhich never saw WAVESEM tokens or reconstructions during trainingdirectly on reconstructed waveforms $\hat{\mathbf{x}}$, alongside simulated

`slow5tools degrade -b` LSB elimination (float → 11-bit ADC → eliminate b LSBs → float) and the frozen 5mC probe:

Table 15: Downstream basecalling accuracy (using an unadapted raw-signal CRF Bonito on reconstructed waveforms $\hat{\mathbf{x}}$) and frozen-encoder 5mC probing AUROC under ex-zd-style LSB degradation versus WAVESEM.

Condition	Raw-Signal Bonito Basecall ACC ↑	Frozen 5mC AUROC ↑
Clean (uncompressed)	0.9719	0.9278
ex-zd `-b 1`	0.9719	0.9273
ex-zd `-b 2`	0.9719	0.9275
ex-zd `-b 3`	0.9719	0.9276
ex-zd `-b 4`	0.9715	0.9280
WAVESEM ($\hat{\mathbf{x}}$ / frozen probe)	0.8500	0.9280

Mild LSB elimination (`-b 1` to `-b 4`) preserves both raw-signal basecalling (0.9715–0.9719) and 5mC probing (0.9273–0.9280) almost exactly. For WAVESEM, feeding the $14.2\times$ reconstructed waveform $\hat{\mathbf{x}}$ into the unadapted raw-signal Bonito model yields a read accuracy of 0.8500 (vs. 0.9719 on clean signal), whereas training the token→CRF decoder directly on WAVESEM tokens reaches 0.9561 at $64\times$ (Table 1). This comparison shows that 0.8500 reflects the standalone waveform fidelity limit of $14.2\times$ lossy reconstruction when consumed by an unadapted raw-signal caller, while the token-adapted decoder compensates for part of the compression loss by operating directly on the disentangled discrete representation. Accordingly, lossless SLOW5/ex-zd is preferable for maximal fidelity and high-stakes archival, lossy ex-zd is suited to modest compression ($\sim$3–5$\times$) with drop-in compatibility for unmodified raw-signal tools, and WAVESEM is suited to aggressive storage reduction ($\sim$14$\times$) paired with token-based re-analysis.

I Reduction Metrics and True Bitrates

I.1 Definition of Reduction Metrics

Throughout the paper, we report reduction factors such as "32×" or "64×", as well as the overall reduction ratio (e.g., "14×"). Here we explicitly state that these quantities refer to the *reduction ratio of time downsampling and storage size*.

All "Œ" factors reported in the main text correspond to the down-sampling rate at this time or the reduced size of the resulting tag storage. Let $x \in \mathbb{R}^T$ represent an original waveform segment of length T. After encoding and quantization, the signal is represented as a sequence of N discrete latent tokens, where $N = T/r$, and r denotes the time downsampling ratio. We define the time subsampling rate as:

$$r = \frac{T}{N}. \tag{20}$$

J Limitations and Future Work

While WAVESEM demonstrates strong empirical performance, several limitations and scope boundaries remain:

- **Workflow role and complementarity.** WAVESEM is designed for two specific roles: (1) lossy archival of compact dual-stream RVQ tokens ($\sim$14.2$\times$ smaller than Int16 waveforms), and (2) a reusable discrete input representation for downstream models trained on frozen WAVESEM features. Although the trained tokenizer encodes new signals out-of-the-box without per-sample optimization, downstream predictors still require task- and domain-specific training or calibration. WAVESEM is therefore complementary to raw-signal formats and production callers such as Dorado and Remorait is not a lossless replacement for SLOW5/BLOW5, not a drop-in replacement for raw-signal Dorado, and not a justification for discarding raw waveforms in high-stakes clinical workflows.
- **Pore chemistry, batch effects, and modality scope.** All experiments are conducted on ONT nanopore sequencing signals. While the frozen R10.4.1 tokenizer transfers zero-shot

across species and GC content with minor degradation, cross-chemistry transfer to R9.4.1 incurs noticeable reconstruction degradation, and nanopore run-to-run or flow-cell batch shifts mean that downstream modification probes should be interpreted via relative information retention rather than uncalibrated clinical accuracy. Generalization to other scientific sensing modalities and adaptation of the fixed Daubechies-4 basis and bit allocation (N_{sem}=6, N_{tex}=2) require further validation.

- **Downstream task coverage and sequence length.** Downstream evaluation covers token-adapted basecalling, enzymatic 5mC CpG probing (segment-level and per-CpG), and weak-label probing of endogenous *E. coli* modifications (dam-6mA and dcm-5mC); tasks such as production site-level 6mA calling, structural variant calling, and poly(A) tail estimation remain future work. Finally, the Transformer in the semantic stream may incur high memory costs on ultra-long reads ($>$1M samples).

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction (Section 1) clearly state three contributionsspectral inductive bias, dual-stream RVQ, and physically consistent self-supervisionall of which are substantiated by experimental results in Sections 4–5. Claims regarding $14\times$ compression, 95.61% basecalling accuracy, and AUROC = 0.928 are directly supported by Tables 1–2 and Figures 2–4.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: A dedicated "Limitations and Future Work" section is provided in Appendix J, discussing domain specificity (ONT-only evaluation), the fixed wavelet basis, fixed bit allocation, limited downstream task coverage, scalability to ultra-long reads, and the absence of formal theoretical guarantees.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [N/A]

   Justification: The paper does not present formal theoretical results, theorems, or complete proofs. Appendix A provides motivational analysis grounded in signal-processing principles (energy conservation, Parseval's equality, and sub-band alignment), and it is explicitly framed as engineering-level motivation rather than formal proof.

   Guidelines:

   - The answer [N/A] means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: The paper provides detailed experimental setup in Appendix D (datasets, preprocessing, baselines) and an Implementation Details appendix section (optimizer, learning rate schedule, batch size, hardware). Model hyperparameters (latent dimension, codebook size, Transformer configuration, DWT settings) and evaluation metrics are also specified in the appendix.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.

(c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).

(d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: The source code and launch scripts for pretraining WAVESEM and baselines, downstream token-based basecalling, 5mC methylation probing, and experimental analyses are openly available at `https://github.com/GEHAH/WaveSem`. The datasets used in the experiments are publicly available and described in Appendix D.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: Training details are provided in the Implementation Details appendix section (AdamW optimizer, cosine annealing schedule, batch size 2,048, 8$\times$ NVIDIA H20 GPUs). Data splits are described in Appendix D (30M self-supervised segments, separate labeled basecalling set, 100K balanced methylation segments with 15% validation split). Model hyperparameters are listed in the appendix.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [No]

   Justification: The main basecalling table reports $\pm$ values, but not all experimental tables include statistical uncertainty.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
   - For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
   - If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Appendix E specifies the hardware (8$\times$ NVIDIA H20 GPUs) and training configuration (batch size 2,048, cosine annealing schedule). Inference cost is analyzed in terms of GFLOPs in Figure 4(c).

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: This work uses publicly available genomic sequencing datasets and proposes a signal compression method. It does not involve human subjects, private data, or dual-use concerns. The research conforms with the NeurIPS Code of Ethics.

   Guidelines:

- The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [No]

    Justification: The paper does not include a dedicated broader-impact discussion. The work is foundational research on signal representation learning for genomic data, with likely positive impact through reduced storage costs and improved accessibility of genomic analysis; no direct negative societal impacts are foreseen.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: The proposed model is a domain-specific signal compressor for genomic data and does not pose risks of misuse comparable to generative language or image models.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.

- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [No]

    Justification: The paper cites the datasets and baseline methods used in the experiments, including HG002 nanopore sequencing, the ONT E. coli 5mC benchmark, Bonito, SoundStream, VQ-VAE, SLOW5/BLOW5, ex-zd, and DeepMod2. However, explicit license names, versions, and URLs are not fully documented in the current manuscript.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [N/A]

    Justification: The paper does not release new datasets or pretrained model checkpoints at submission time. Code and models will be released upon acceptance.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: This work does not involve crowdsourcing or research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.

- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: This work does not involve new human-subject recruitment, intervention, or data collection. The genomic data used (HG002, E. coli) are publicly available reference datasets.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A]

    Justification: The core methodology does not involve LLMs as any important, original, or non-standard component.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.