SPANDNA: RESOLUTION-AWARE SPECTRAL SUPERVISION FOR GENOMIC PRETRAINING

Anonymous authors
Paper under double-blind review

ABSTRACT

Masked language modeling (MLM) does not explicitly specify which sequence-wide statistics should be readily accessible from a pooled genomic representation. We study a training-only auxiliary objective based on the *Resolution-aware Nucleotide Spectrum* (RNS), a compact Fourier representation of nucleotide indicator signals that combines composition with phase-sensitive low-frequency coordinates. SpanDNA predicts a 36-dimensional RNS-4 target jointly with MLM and discards the auxiliary head after pretraining. Using a 117M Transformer with the same pretraining data and update budget, three pretraining seeds, and three finetuning seeds per checkpoint, mean MCC across 18 tasks increases from 0.6421 to 0.6500. The mean gain is positive for each of the three reported pretraining seeds, and 15 of 18 task-level means improve. Fixed-descriptor experiments show that RNS contains predictive signal beyond composition under linear classification. In a single-pair representation analysis, frozen linear probes recover the 32 non-DC coordinates with $R^2 = 0.909$ after RNS supervision, compared with -0.176 under MLM alone. These results show that RNS is an effective auxiliary target in the evaluated setting and that it changes the linear accessibility of supervised sequence-level statistics without modifying the deployed encoder.

1 INTRODUCTION

Genomic language models have emerged as a general-purpose approach for learning transferable representations from large collections of unlabeled DNA sequences (Ji et al., 2021; Dalla-Torre et al., 2025; Zhou et al., 2024). Most models rely on masked or autoregressive token prediction during pretraining and are subsequently adapted to regulatory annotation, chromatin profiling, splice-site prediction, and related tasks. Recent evaluations further show that downstream transfer can depend strongly on tokenization and architecture, while gains attributable to pretraining are not always large (Vishniakov et al., 2026). An additional design question concerns sequence-level readout: token prediction provides rich contextual supervision, but many downstream tasks classify or score an entire sequence window from a pooled representation. The issue is not whether MLM can encode sequence-wide information, but that its objective does not directly specify which whole-sequence statistics should be readily accessible from the representation used by downstream classifiers.

DNA offers a natural opportunity for such supervision because global properties can be computed directly from the input without external annotations. Nucleotide composition is the simplest example, but it discards arrangement: `AAAACCCCGGGGTTTT` and `ACGTACGTACGTACGT` have identical base counts but very different positional organization. More generally, equal-composition sequences can differ in block structure, positional bias, and periodicity. A useful analytical target can therefore complement token prediction by explicitly shaping access to compact sequence-level statistics while remaining annotation-free.

We formulate the *Resolution-aware Nucleotide Spectrum* (RNS), which represents nucleotide indicator signals in the Fourier domain and retains coefficients up to a controllable order R. At $R = 0$, RNS reduces to nucleotide composition; increasing R adds progressively finer position-sensitive coordinates, while the complete nonnegative-frequency spectrum determines the original sequence. SpanDNA uses a compact RNS-4 target as training-only supervision: a pooled sequence representation is optimized jointly for masked language modeling and RNS prediction, and the auxiliary head

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

is removed after pretraining. Here, "resolution-aware" refers to the target family indexed by R; all reported SpanDNA pretraining runs use the fixed choice $R = 4$.

Across three pretraining repetitions per objective, mean MCC over 18 tasks increases from 0.6421 to 0.6500; all three pretraining-seed gains are positive and 15 of 18 task-level means improve. Fixed-descriptor and frozen-probe analyses separately characterize the target's predictive signal and its linear accessibility from the encoder. We therefore evaluate whether this supervision improves transfer through repeated pretraining comparisons while keeping the deployed encoder unchanged.

Our contributions are:

- We formulate RNS as a compact, annotation-free sequence-level target whose resolution ranges from nucleotide composition toward the full spectrum, and characterize its invertibility, nested spectral distances, and full-resolution relation to Hamming distance.
- We develop SpanDNA, which predicts RNS jointly with MLM during genomic pretraining while leaving the downstream encoder architecture and parameter count unchanged.
- We demonstrate a positive mean downstream effect across three pretraining repetitions on 18 tasks and separately characterize the increased linear recoverability of supervised non-DC statistics from frozen pooled representations.

2 Related Work

Genomic sequence models and foundation models. Sequence-to-function models such as DeepSEA, Basset, and Enformer showed that neural networks can learn regulatory activity directly from DNA and benefit from broader sequence context (Zhou & Troyanskaya, 2015; Kelley et al., 2016; Avsec et al., 2021). Genomic foundation models instead use large-scale self-supervised pretraining, commonly adapting Transformer/BERT-style modeling and text-inspired tokenization (Vaswani et al., 2017; Devlin et al., 2019; Sennrich et al., 2016). DNABERT uses fixed k-mers (Ji et al., 2021), whereas DNABERT-2 and GROVER use byte-pair encoding (Zhou et al., 2024; Sanabria et al., 2024). Nucleotide Transformer studies scaling across human and multi-species genomes (Dalla-Torre et al., 2025); GENA-LM extends masked modeling to longer contexts (Fishman et al., 2025); HyenaDNA and Caduceus explore long-convolution and bidirectional state-space architectures (Nguyen et al., 2023; Schiff et al., 2024); GPN supports genome-wide variant-effect prediction (Benegas et al., 2023); and Evo/Evo 2 extend autoregressive modeling toward genome-scale context and design (Nguyen et al., 2024; Brixi et al., 2026). Tokenization is also an active design axis, with learned tokenization in MxDNA (Qiao et al., 2024), biologically informed patching in PatchDNA (Del Vecchio et al., 2026), and dynamic chunking in DNAChunker (Kim et al., 2026). SpanDNA changes a different axis: the deployed encoder and tokenizer remain conventional, while pretraining explicitly constrains what the pooled representation preserves.

Objectives that shape biological representations. Beyond masked token reconstruction, auxiliary and multi-task constraints can shape learned representations (Caruana, 1997). DNAGPT adds sequence-order classification and GC-content regression (Zhang et al., 2023); species-aware DNA language models incorporate organism identity to capture conserved regulatory structure (Gankin et al., 2024); DNABERT-S uses contrastive, mixup-based objectives for species-aware embeddings (Zhou et al., 2025); and LucaOne combines nucleic-acid/protein modeling with annotation-based objectives (He et al., 2025). Like composition-based auxiliary objectives, RNS is computed directly from the input without external annotations. Its distinguishing design choice is to supervise a resolution-indexed set of phase-sensitive nucleotide coordinates from the pooled representation, extending composition toward position-sensitive organization rather than introducing an external label source or downstream feature.

Alignment-free and spectral sequence representations. Alignment-free methods summarize sequences without explicit pairwise alignment (Zielezinski et al., 2017), including k-mer and gapped-k-mer features (Lee et al., 2011; Ghandi et al., 2014; Lee, 2016), Chaos Game Representation (Jeffrey, 1990), and compact MinHash sketches such as Mash (Ondov et al., 2016). Fourier analysis provides a complementary representation: early work showed that spectral magnitude and phase can encode biologically useful sequence structure (Anastassiou, 2000), while later methods used power

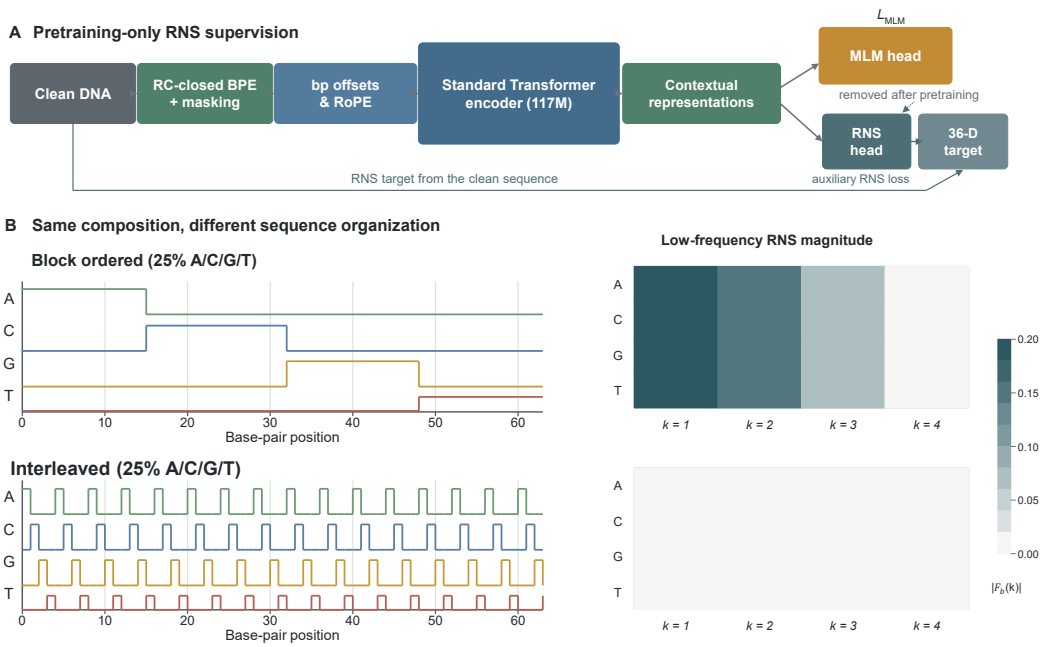


Figure 1: Overview of SpanDNA. A reverse-complement-closed BPE tokenizer maps DNA into variable-span tokens, which are processed by a standard Transformer encoder. During pretraining, a base-pair-span-weighted sequence representation predicts the 36-dimensional RNS-4 target in addition to masked tokens. The RNS head is removed after pretraining. The lower panel visualizes equal-composition sequences using magnitude only; the training target retains real and imaginary coordinates.

or magnitude spectra for sequence comparison and classification (Randhawa et al., 2019). Neural genomics has also used Fourier structure as attribution priors (Tseng et al., 2020), reconstruction regularization (Patel & Kundaje, 2026), and position embeddings (Wang et al., 2026). RNS retains real and imaginary low-frequency coefficients of nucleotide indicator signals and uses them only as a training target. Thus phase is preserved, while neither spectral features nor the auxiliary head are required at inference time. Building on spectral sequence representations and auxiliary pretraining, SpanDNA studies a compact, phase-retaining nucleotide spectrum as explicit supervision for pooled genomic representations. We evaluate this training-only target through repeated downstream comparisons and frozen-representation diagnostics.

3 Method

SpanDNA introduces sequence-level RNS supervision into masked genomic pretraining while leaving the encoder architecture unchanged. As illustrated in Figure 1, the nucleotide sequence is used in two parallel ways during pretraining: the tokenized sequence is processed by the Transformer for masked language modeling, while the same sequence is converted into a Resolution-aware Nucleotide Spectrum (RNS) target. A sequence-level encoder representation is trained to predict this target jointly with masked tokens. RNS is used only as a training target and is not appended to the model input; its prediction head is removed after pretraining.

3.1 Resolution-aware Nucleotide Spectrum

Let $S = (s_0, \ldots, s_{N-1})$ be a DNA sequence of length N over $\mathcal{B} = \{A, C, G, T\}$. During training, S is the clean sequence span visible to the tokenizer after reverse-complement augmentation and truncation; MLM corruption is applied only after this target sequence is fixed. For each nucleotide $b \in \mathcal{B}$, we represent its occurrence along the sequence by a binary indicator signal

$$x_b(t) = \mathbf{1}\{s_t = b\}, \qquad t = 0, \ldots, N-1, \tag{1}$$

where $\mathbf{1}\{\cdot\}$ denotes the indicator function.

Each $x_b(t)$ is a discrete signal indexed by base-pair position. Its Fourier representation decomposes nucleotide organization into components at different spatial scales: the zero-frequency component measures overall abundance, low frequencies describe coarse variation across the sequence window, and progressively higher frequencies represent finer positional variation. We therefore define the normalized discrete Fourier coefficients as

$$F_b(k) = \frac{1}{N} \sum_{t=0}^{N-1} x_b(t) \exp\left(\frac{2\pi i k t}{N}\right), \tag{2}$$

where $i = \sqrt{-1}$.

The coefficient $F_b(0) = \frac{1}{N}\sum_{t=0}^{N-1} x_b(t)$ is exactly the fraction of nucleotide b in the sequence. For a maximum frequency order $R \leq \lfloor N/2 \rfloor$, the Resolution-aware Nucleotide Spectrum is obtained by concatenating the zero-frequency coefficient with the real and imaginary coordinates of the first R nonzero frequencies:

$$\Phi_R(S) = \mathop{\Big\|}_{b\in\mathcal{B}} \left[F_b(0), \operatorname{Re} F_b(1), \operatorname{Im} F_b(1), \ldots, \operatorname{Re} F_b(R), \operatorname{Im} F_b(R)\right]. \tag{3}$$

For the truncated resolutions considered in this work, its dimension is $4(1 + 2R)$. Retaining both real and imaginary coordinates preserves Fourier phase and therefore the location of spectral patterns relative to the sequence origin, information that is lost when only spectral magnitude is retained. The phase reference is the beginning of the visible sequence window; RNS is not intended to be translation invariant.

Because the frequency index k is defined relative to normalized position t/N, the same order R describes organization at a relative spatial scale across sequence lengths. At $R = 0$, RNS reduces exactly to nucleotide composition. Increasing R adds progressively finer position-sensitive information about how nucleotides are arranged along the sequence. The following results characterize the limiting case in which all nonnegative frequencies required by conjugate symmetry are retained.

3.2 Theoretical Properties of RNS

Proposition 1 (Completeness). For a canonical DNA sequence of fixed length N, let

$$R_\star = \left\lfloor \frac{N}{2} \right\rfloor.$$

The full nonnegative-frequency RNS at $R_\star$ uniquely determines the original sequence.

Each indicator signal $x_b(t)$ is real-valued, and its Fourier coefficients therefore satisfy conjugate symmetry,

$$F_b(N-k) = \overline{F_b(k)}.$$

Thus the retained coefficients determine the omitted negative frequencies; inverse DFT then recovers the four indicator signals and hence the sequence. The representation is complete but not coordinate-minimal because of the Nyquist and indicator-sum constraints. A complete proof is given in Appendix A.

Proposition 2 (Spectral–Hamming identity). For two canonical sequences S and T of equal length N, define the truncated weighted spectral distance

$$D_R^2(S,T) = \sum_{b\in\mathcal{B}} \sum_{k=0}^{R} w_k \left|F_b^S(k) - F_b^T(k)\right|^2, \tag{4}$$

where $w_0 = 1$, conjugate frequency pairs have weight $w_k = 2$, and the Nyquist term has weight $w_{N/2} = 1$ when N is even. At full nonnegative-frequency resolution,

$$D_{R_\star}^2(S,T) = \frac{2}{N} d_H(S,T), \tag{5}$$

where $d_H(S,T)$ denotes Hamming distance. Moreover,

$$D_0^2(S,T) \le D_1^2(S,T) \le \cdots \le D_{R_\star}^2(S,T). \tag{6}$$

By Parseval's theorem, each nucleotide mismatch contributes squared indicator difference 2, yielding Eq. 5; increasing R only adds nonnegative spectral energy, giving Eq. 6. The full derivation is in Appendix A.

For $R < R_\star$, RNS is a resolution-controlled projection and need not permit exact reconstruction; the Hamming identity applies only at full weighted resolution. These properties characterize how the RNS family interpolates between composition and complete sequence information. We use RNS-4 as a compact low-frequency target and evaluate its utility empirically.

Reverse-complement transformation. RNS also transforms deterministically under reverse complementation. Let $\bar{b}$ denote the complement of nucleotide b and let S^{RC} be the reverse complement of S. Then

$$F_b^{S^{\mathrm{RC}}}(k) = \exp\left(-\frac{2\pi i k}{N}\right) \overline{F_{\bar{b}}^{S}(k)}. \tag{7}$$

Thus reverse complementation maps RNS through a fixed nucleotide-channel permutation, complex conjugation, and a frequency-dependent phase factor. The derivation is provided in Appendix A.

3.3 RNS-supervised Pretraining

The genomic encoder uses variable-span BPE tokens. We retain each token's base-pair span and evaluate rotary position embeddings at token centers in base-pair coordinates.

Let $H = (h_1, \ldots, h_L)$ denote the final hidden states of the L tokens. Let m_j be the number of base pairs covered by token j and $c_j \in \{0,1\}$ indicate non-special sequence content. Because the RNS target is defined over base-pair positions while BPE tokens cover variable spans, an unweighted token average would make the sequence-level representation depend unnecessarily on the particular segmentation. We instead aggregate variable-length tokens with a base-pair-span-weighted mean, so token contributions reflect their covered base-pair spans:

$$h_S = \frac{\sum_{j=1}^{L} m_j c_j h_j}{\sum_{j=1}^{L} m_j c_j}. \tag{8}$$

A lightweight two-layer prediction head g_ψ maps h_S to an RNS target. We use $R = 4$, which retains the first four nonzero spatial frequencies for each nucleotide and produces a 36-dimensional target,

$$\Phi_4(S) \in \mathbb{R}^{36}.$$

This keeps the auxiliary task focused on a compact low-frequency description. Section 4.2 characterizes how fixed-descriptor predictive information changes with R.

The pretraining objective combines masked language modeling with RNS regression:

$$\mathcal{L} = \mathcal{L}_{\mathrm{MLM}} + \lambda(t)\,\mathrm{SmoothL1}\left(g_\psi(h_S), \Phi_4(S)\right), \tag{9}$$

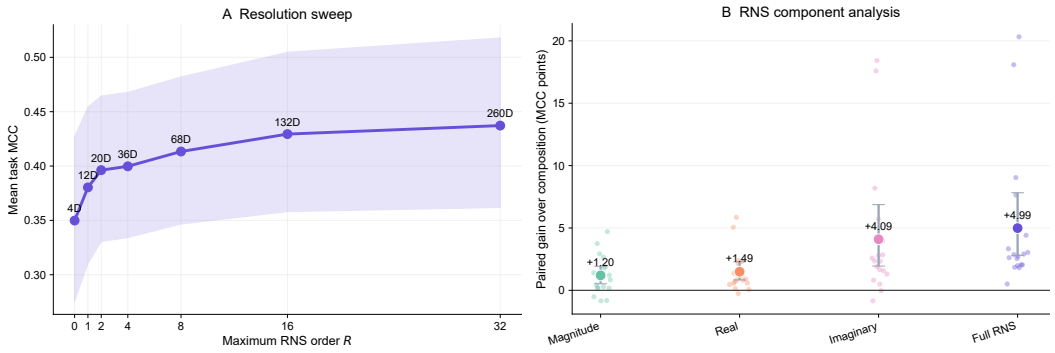
where $\lambda(t)$ controls the RNS loss and is linearly warmed up at the beginning of pretraining. Prediction-head and optimization details are given in Appendix B.

RNS supervision is used only during pretraining. After pretraining, g_ψ is removed and all downstream tasks use the same encoder architecture and parameter count as the MLM-only model.

4 Experiments

4.1 Experimental Setup

Pretraining. We use a 117M-parameter bidirectional Transformer with a reverse-complement-closed BPE tokenizer (3,991 deployed tokens). Pretraining uses 1,000-bp windows, at most 512 tokens, 15% MLM corruption, and reverse-complement augmentation with probability 0.5. The



(a) Resolution sweep.

(b) RNS-4 components.

Figure 2: Predictive information in RNS under fixed-descriptor classification. (a) Mean MCC across 18 tasks as the retained frequency order increases; $R = 0$ is composition. (b) Gain over composition for RNS-4 component subsets; each non-DC subset is evaluated together with the four composition coordinates. Shading and error bars are 95% percentile bootstrap intervals over task-level means (10,000 resamples).

35.6M-window corpus combines the DNABERT-2 multi-species corpus with T2T-CHM13. Each run uses 150,000 optimization steps and an effective batch size of 1,024. For RNS training, the target is computed from the clean visible sequence after reverse-complement augmentation and before MLM corruption. We train three runs for each objective using pretraining seeds 42, 43, and 44. Additional architecture, optimization, and target-construction details are in Appendix B.

Downstream evaluation. We use the 18 Nucleotide Transformer classification tasks (Dalla-Torre et al., 2025): ten histone-mark, two enhancer, three promoter, and three splice-site tasks. Each checkpoint is fully finetuned for three epochs with seeds 41, 42, and 43; 10% of the official training set is held out for validation, and the best epoch is selected by validation MCC. MCC is the primary metric, with macro F1, accuracy, and AUROC reported in the Appendix. For each task and pretraining seed, we first average the three finetuning seeds and then summarize across pretraining seeds. Task-level paired analyses average both seed levels before treating the 18 tasks as paired units. Complete finetuning and statistical protocols are provided in Appendices C and E.

4.2 RNS Captures Predictive Information Beyond Composition

We first test whether nonzero-frequency coordinates add predictive information beyond composition under fixed linear classification. Mean MCC rises from 0.3499 at $R = 0$ to 0.3997 at $R = 4$ and 0.4372 at $R = 32$; RNS-4 improves all 18 tasks over composition while using 36 dimensions versus 260 at $R = 32$ (Figure 2a). We use $R = 4$ as a compact target; this sweep characterizes descriptor information, not a pretraining-optimal order.

At $R = 4$, composition plus spectral magnitude gains 1.20 MCC points over composition, compared with 4.99 for the full real-and-imaginary representation; real-only and imaginary-only gains are 1.49 and 4.09 (Figure 2b). The stronger full representation indicates that phase-sensitive positional information contributes useful predictive signal in the fixed-descriptor setting.

4.3 RNS Supervision Improves Downstream Performance

Across the three pretraining seeds, MLM-only achieves 0.6421 ± 0.0008 mean MCC and RNS-supervised pretraining reaches 0.6500 ± 0.0015 (Table 1b). The mean improvement is $+0.0079 \pm 0.0022$, with positive gains for seeds 42, 43, and 44. After averaging both pretraining and finetuning seeds within each task, RNS improves 15 of 18 task-level means; the mean and median differences are $+0.0079$ and $+0.0070$, with a 95% bootstrap interval of $[+0.0023, +0.0137]$ and two-sided Wilcoxon $p = 0.0104$. Secondary metrics show the same overall direction (Appendix F.2).

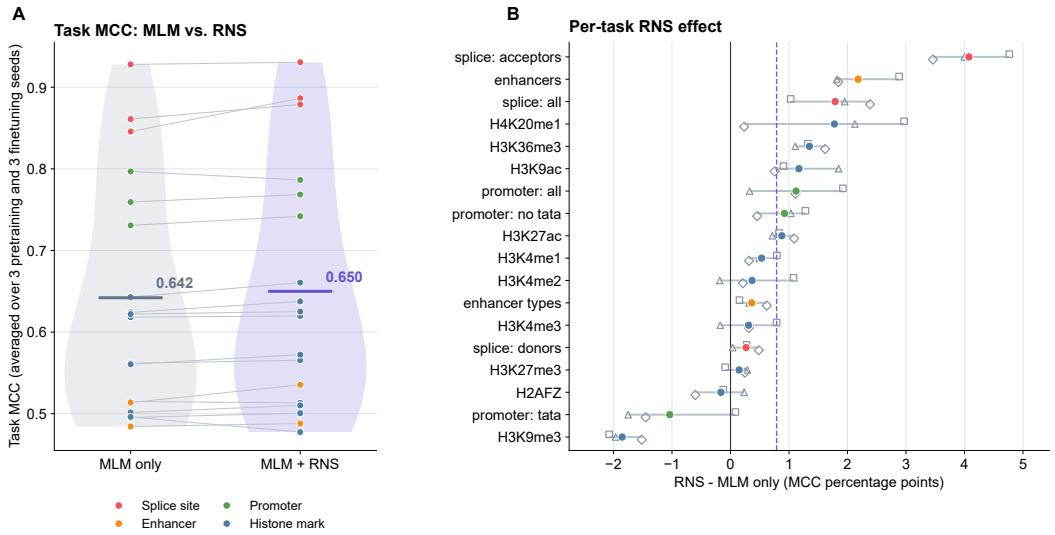


Figure 3: Downstream comparison across pretraining seeds 42, 43, and 44. Left: per-task MCC after averaging pretraining and finetuning seeds. Right: RNS-minus-MLM task differences; symbols show individual pretraining seeds after averaging finetuning seeds, and gray ranges span seed-level effects.

Effects are not uniform: H2AFZ, promoter-TATA, and H3K9me3 have negative mean differences. Positive effects occur across all four task families, and every leave-one-family-out mean remains positive, including $+0.54$ MCC points after excluding splice-site tasks (Appendix F.3).

4.4 Comparison with Existing DNA Foundation Models

Table 1a provides benchmark context. DNABERT-2 (Zhou et al., 2024), ModernGENA (Aspidova et al., 2026), and DNABERT-1 (Ji et al., 2021) are locally rerun with our three-epoch protocol; MxDNA (Qiao et al., 2024), PatchDNA (Del Vecchio et al., 2026), Caduceus (Schiff et al., 2024), GENA-LM (Fishman et al., 2025), HyenaDNA (Nguyen et al., 2023), and phyloP (Pollard et al., 2010) follow the published PatchDNA comparison. The SpanDNA entry reports the mean across pretraining seeds 42, 43, and 44 after averaging three finetuning seeds per checkpoint, with complete per-task results reported in Appendix F.1. Across the three independent pretraining seeds, SpanDNA achieves a mean MCC of 0.6500, with seed-level means ranging from 0.6487 to 0.6516 (Table 1b). Because evaluation budgets differ across published and locally rerun models, panel (a) is provided as benchmark context.

4.5 RNS Supervision Changes Learned Representations

We evaluate target prediction and frozen linear readout separately. On 2,000 H3K4me1 test sequences, the RNS head predicts the 32 non-DC coordinates with variance-weighted $R^2 = 0.9361$, Pearson $r = 0.9682$, Spearman $\rho = 0.9614$, and MAE 0.0030. Analytical and predicted similarity maps correlate at $r = 0.9177$ on 256 sequences (Figure 4A–D).

Using the seed-42 models from Table 1b, frozen ridge probes fitted to 5,000 H3K4me1 sequences and tested on 2,000 disjoint sequences achieve final-layer non-DC R^2 of -0.1760 (MLM) and 0.9089 (RNS), with separation mainly in later layers (Figure 4E). DC composition is highly recoverable from both (0.9869 vs. 0.9936). The supervised statistics thus remain accessible through a linear readout after removing the nonlinear auxiliary head, with the largest separation emerging in the later encoder layers (Appendix D).

4.6 Auxiliary-Target Comparisons

Table 1c compares additional auxiliary targets under the same architecture, data configuration, optimization budget, and downstream protocol. Composition reaches 0.6288 MCC, Global RNS

Table 1: Downstream evaluation of SpanDNA. (a) Benchmark context; the SpanDNA result is averaged across pretraining seeds 42, 43, and 44 after averaging the three finetuning seeds for each checkpoint. (b) Results across the three pretraining seeds; mean±SD is computed across pretraining seeds. (c) Auxiliary-target comparisons using pretraining seed 42; ΔMCC is relative to MLM only in percentage points. *Locally rerun with three finetuning seeds. †Published values reported in the PatchDNA comparison. Complete per-task results are in the Appendix.

(a) Benchmark context

Model	Publication	Mean MCC
SpanDNA	Ours	0.6500
DNABERT-2 (117M)*	ICLR 2024	0.6496
ModernGENA-Base*	ICLR 2026	0.6392
MxDNA†	NeurIPS 2024	0.6374
PatchDNA†	ICLR 2026	0.6257
DNABERT-1 (6-mer)*	Bioinform. 2021	0.6228
PatchDNA Entropy†	ICLR 2026	0.6175
Caduceus-ps†	ICML 2024	0.6118
GENA-LM-Base†	NAR 2025	0.5995
HyenaDNA†	NeurIPS 2023	0.5986
PhyloP†	Genome Res. 2010	0.2092

(b) Three-seed pretraining comparison

Seed	MLM	+RNS	Δ
42	0.6412	0.6516	+0.0104
43	0.6428	0.6496	+0.0068
44	0.6423	0.6487	+0.0064
Mean	0.6421±0.0008	0.6500±0.0015	0.0079±0.0022

(c) Auxiliary-target comparison

Target	Dim.	MCC	ΔMCC
MLM only	–	0.6412	0.00
Composition	4	0.6288	−1.24
Global RNS	36	0.6516	+1.04
Multiscale RNS	252	0.6519	+1.07

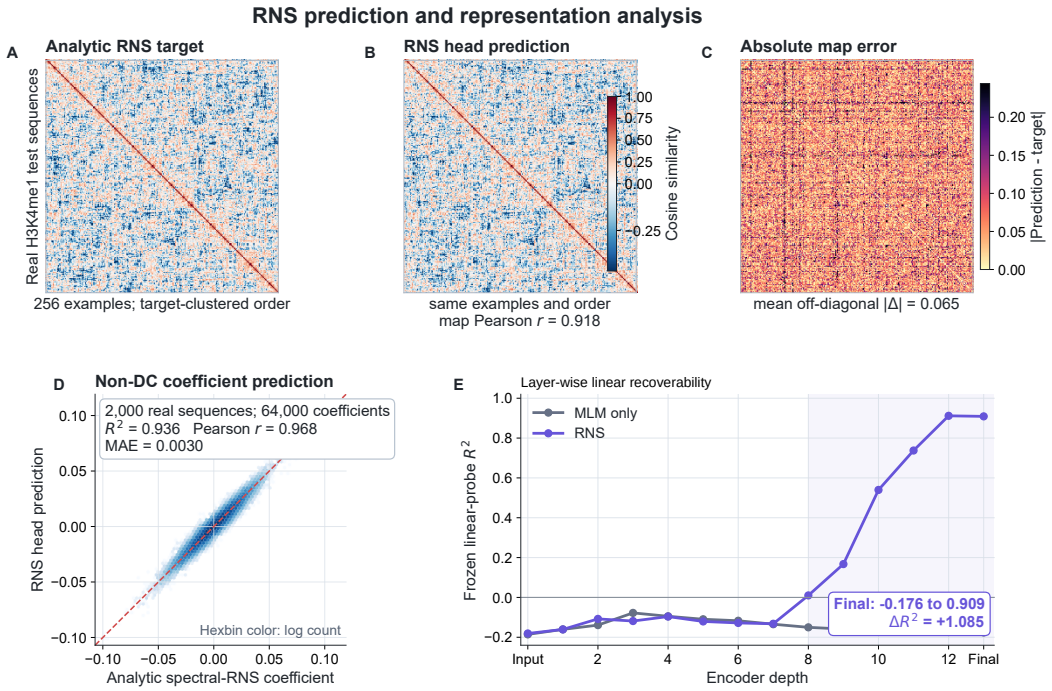


Figure 4: Representation analysis. Panels A–C compare analytical and RNS-head similarity maps on 256 held-out H3K4me1 sequences. Panel D shows 64,000 non-DC coordinate predictions from 2,000 sequences; these coordinates are not treated as independent sequence-level samples. Panel E shows layer-wise frozen ridge probes trained on 5,000 sequences and evaluated on 2,000 disjoint sequences for the seed-42 MLM/RNS pair. RNS supervision produces a pronounced late-layer increase in linear recoverability of the supervised nonzero-frequency coordinates.

reaches 0.6516, and Multiscale RNS reaches 0.6519. Global and Multiscale RNS are nearly identical in this comparison (0.6516 vs. 0.6519), while the composition-only variant is lower (0.6288). Training diagnostics for these auxiliary-target runs are provided in Appendix G.1. In this seed-42 comparison, the gain is not reproduced by composition-only supervision, but these experiments do not determine whether the benefit is specific to the spectral parameterization or extends to other position-aware sequence-level targets.

After pretraining, the RNS head is removed, leaving the same 116,352,770-parameter encoder with no additional downstream inference cost.

5 Discussion and Limitations

SpanDNA shows that explicit supervision of analytically defined sequence-wide statistics can improve average downstream performance without changing the deployed encoder. Across three pretraining seeds, RNS supervision yields a consistent positive mean effect, while the descriptor and probing analyses show that the target contains task-relevant signal and becomes substantially more linearly recoverable from pooled representations. These results support RNS as a compact training-only objective for shaping genomic sequence representations.

The present study is limited to one 117M architecture, one pretraining scale, sequence windows of at most 1,000 bp, and a three-epoch downstream finetuning protocol. Effects vary across tasks. The fixed-descriptor resolution sweep characterizes the target family but does not determine the best pretraining order. Broader comparisons with alternative position-aware sequence summaries would further clarify which gains are specific to the spectral parameterization and which arise more generally from sequence-level auxiliary supervision. Evaluation at larger model scales, longer sequence contexts, and broader downstream settings remains open.

AI Use Statement

Generative AI tools were used to assist with language polishing, literature retrieval and discovery, and feedback on manuscript organization and the interpretation of existing results. They were not used to generate experimental data or determine reported numerical results. All AI-assisted content was reviewed and verified by the authors, who take full responsibility for the final manuscript.

Reproducibility Statement

The paper and appendices document the experimental procedures used in this study. Appendix B specifies the tokenizer, RNS target construction, model architecture, pretraining objective, optimization settings, random seeds, and hardware configuration. Appendices C–E describe downstream evaluation, representation probing, and statistical reporting, and Appendix G reports auxiliary-target training diagnostics and matched inference measurements. Anonymous source code, the frozen tokenizer, and configuration files for the core MLM-only and RNS pretraining experiments are provided in the supplementary material.

References

Dimitris Anastassiou. Frequency-domain analysis of biomolecular sequences. *Bioinformatics*, 16 (12):1073–1081, 2000. doi: 10.1093/bioinformatics/16.12.1073.

Alena Aspidova, Yuri Kuratov, Artem Shadskiy, Mikhail Burtsev, and Veniamin Fishman. Back to BERT in 2026: ModernGENA as a strong, efficient baseline for DNA foundation models. In *International Conference on Learning Representations*, 2026.

Žiga Avsec, Vikram Agarwal, Daniel Visentin, Joseph R. Ledsam, Agnieszka Grabska-Barwinska, Kyle R. Taylor, Yannis Assael, John Jumper, Pushmeet Kohli, and David R. Kelley. Effective gene expression prediction from sequence by integrating long-range interactions. *Nature Methods*, 18 (10):1196–1203, 2021. doi: 10.1038/s41592-021-01252-x.

Gonzalo Benegas, Sanjit Singh Batra, and Yun S. Song. DNA language models are powerful predictors of genome-wide variant effects. *Proceedings of the National Academy of Sciences*, 120 (44):e2311219120, 2023. doi: 10.1073/pnas.2311219120.

Garyk Brixi, Matthew G. Durrant, Jerome Ku, Mohsen Naghipourfar, Michael Poli, Gwanggyu Sun, Greg Brockman, Daniel Chang, Gabriel A. Gonzalez, Samuel H. King, et al. Genome modelling and design across all domains of life with Evo 2. *Nature*, 652:1349–1361, 2026. doi: 10.1038/s41586-026-10176-5.

Rich Caruana. Multitask learning. *Machine Learning*, 28:41–75, 1997.

Hugo Dalla-Torre, Liam Gonzalez, Javier Mendoza-Revilla, Nicolas Lopez Carranza, Adam Henryk Grzywaczewski, Francesco Oteri, Christian Dallago, Evan Trop, Bernardo P. de Almeida, Hassan Sirelkhatim, et al. Nucleotide transformer: Building and evaluating robust foundation models for human genomics. *Nature Methods*, 22:287–297, 2025. doi: 10.1038/s41592-024-02523-z.

Alice Del Vecchio, Chantriolnt-Andreas Kapourani, Abdullah M. Athar, Agnieszka Dobrowolska, Andrew Anighoro, Benjamin Tenmann, Lindsay Edwards, and Cristian Regep. PatchDNA: A flexible and biologically-informed alternative to tokenization for DNA. In *International Conference on Learning Representations*, 2026.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*, 2019.

Veniamin Fishman, Yuri Kuratov, Aleksei Shmelev, Maxim Petrov, Dmitry Penzar, Denis Shepelin, Nikolay Chekanov, Olga Kardymon, and Mikhail Burtsev. GENA-LM: A family of open-source foundational DNA language models for long sequences. *Nucleic Acids Research*, 53(2):gkae1310, 2025. doi: 10.1093/nar/gkae1310.

Dennis Gankin, Alexander Karollus, Martin Grosshauser, Kristian Klemon, Johannes Hingerl, and Julien Gagneur. Species-aware DNA language models capture regulatory elements and their evolution. *Genome Biology*, 25:83, 2024. doi: 10.1186/s13059-024-03221-x.

Mahmoud Ghandi, Dongwon Lee, Morteza Mohammad-Noori, and Michael A. Beer. Enhanced regulatory sequence prediction using gapped k-mer features. *PLOS Computational Biology*, 10(7): e1003711, 2014. doi: 10.1371/journal.pcbi.1003711.

Yong He, Pan Fang, Yongtao Shan, Yuanfei Pan, Yanhong Wei, Yichang Chen, Yihao Chen, Yi Liu, Zhenyu Zeng, Zhan Zhou, et al. Generalized biological foundation model with unified nucleic acid and protein language. *Nature Machine Intelligence*, 7:942–953, 2025. doi: 10.1038/s42256-025-01044-4.

H. Joel Jeffrey. Chaos game representation of gene structure. *Nucleic Acids Research*, 18(8): 2163–2170, 1990. doi: 10.1093/nar/18.8.2163.

Yanrong Ji, Zhihan Zhou, Han Liu, and Ramana V. Davuluri. DNABERT: Pre-trained bidirectional encoder representations from transformers model for DNA-language in genome. *Bioinformatics*, 37(15):2112–2120, 2021. doi: 10.1093/bioinformatics/btab083.

David R. Kelley, Jasper Snoek, and John L. Rinn. Basset: Learning the regulatory code of the accessible genome with deep convolutional neural networks. *Genome Research*, 26(7):990–999, 2016. doi: 10.1101/gr.200535.115.

Taewon Kim, Jihwan Shin, Hyomin Kim, Youngmok Jung, Jonhoon Lee, Won-Chul Lee, Insu Han, and Sungsoo Ahn. DNAChunker: Learnable tokenization for DNA language models. *arXiv preprint arXiv:2601.03019*, 2026.

Dongwon Lee. LS-GKM: A new gkm-SVM for large-scale datasets. *Bioinformatics*, 32(14): 2196–2198, 2016. doi: 10.1093/bioinformatics/btw142.

Dongwon Lee, Rachel Karchin, and Michael A. Beer. Discriminative prediction of mammalian enhancers from dna sequence. *Genome Research*, 21(12):2167–2180, 2011. doi: 10.1101/gr.121905.111.

Eric Nguyen, Michael Poli, Marjan Faizi, Armin Thomas, Callum Birch-Sykes, Michael Wornow, Aman Patel, Clayton Rabideau, Stefano Massaroli, Yoshua Bengio, Stefano Ermon, Stephen A. Baccus, and Christopher Ré. HyenaDNA: Long-range genomic sequence modeling at single nucleotide resolution. In *Advances in Neural Information Processing Systems*, 2023.

Eric Nguyen, Michael Poli, Matthew G. Durrant, Brian Kang, Dhruva Katrekar, David B. Li, Liam J. Bartie, Armin W. Thomas, Samuel H. King, Garyk Brixi, et al. Sequence modeling and design from molecular to genome scale with Evo. *Science*, 386(6723):eado9336, 2024. doi: 10.1126/science.ado9336.

Brian D. Ondov, Todd J. Treangen, Páll Melsted, Adam B. Mallonee, Nicholas H. Bergman, Sergey Koren, and Adam M. Phillippy. Mash: Fast genome and metagenome distance estimation using MinHash. *Genome Biology*, 17:132, 2016. doi: 10.1186/s13059-016-0997-x.

Aman Patel and Anshul Kundaje. ARSENAL: Learning transferable regulatory DNA representations with targeted short-context language models. *bioRxiv*, 2026. doi: 10.64898/2026.02.05.703637.

Katherine S. Pollard, Melissa J. Hubisz, Kate R. Rosenbloom, and Adam Siepel. Detection of nonneutral substitution rates on mammalian phylogenies. *Genome Research*, 20(1):110–121, 2010. doi: 10.1101/gr.097857.109.

Lifeng Qiao, Peng Ye, Yuchen Ren, Weiqiang Bai, Chaoqi Liang, Xinzhu Ma, Nanqing Dong, and Wanli Ouyang. Model decides how to tokenize: Adaptive DNA sequence tokenization with MxDNA. In *Advances in Neural Information Processing Systems*, volume 37, 2024.

Gurjit S. Randhawa, Kathleen A. Hill, and Lila Kari. ML-DSP: Machine learning with digital signal processing for ultrafast, accurate, and scalable genome classification at all taxonomic levels. *BMC Genomics*, 20:267, 2019. doi: 10.1186/s12864-019-5571-y.

Melissa Sanabria, Jonas Hirsch, Pierre M. Joubert, et al. DNA language model GROVER learns sequence context in the human genome. *Nature Machine Intelligence*, 6:911–923, 2024. doi: 10.1038/s42256-024-00872-0.

Yair Schiff, Chia-Hsiang Kao, Aaron Gokaslan, Tri Dao, Albert Gu, and Volodymyr Kuleshov. Caduceus: Bi-directional equivariant long-range DNA sequence modeling. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, 2024.

Rico Sennrich, Barry Haddow, and Alexandra Birch. Neural machine translation of rare words with subword units. In *Proceedings of ACL*, 2016.

Alex M. Tseng, Avanti Shrikumar, and Anshul Kundaje. Fourier-transform-based attribution priors improve the interpretability and stability of deep learning models for genomics. In *Advances in Neural Information Processing Systems*, volume 33, 2020.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, 2017.

Kirill Vishniakov, Karthik Viswanathan, Aleksandr Medvedev, Praveenkumar Kanithi, Marco Pimentel, Ronnie Rajan, and Shadab Khan. Tokenization to transfer: Do genomic foundation models learn good representations? In *International Conference on Learning Representations*, 2026.

Weihua Wang, Haoji Li, Feilong Bao, Lei Yang, and Guanglai Gao. Wisteria: A unified multi-scale feature learning framework for DNA language model. *Pattern Recognition*, 180:114289, 2026. doi: 10.1016/j.patcog.2026.114289.

Daoan Zhang, Weitong Zhang, Yu Zhao, Jianguo Zhang, Bing He, Chenchen Qin, and Jianhua Yao. DNAGPT: A generalized pre-trained tool for versatile DNA sequence analysis tasks. *arXiv preprint arXiv:2307.05628*, 2023.

Jian Zhou and Olga G. Troyanskaya. Predicting effects of noncoding variants with deep learning-based sequence model. *Nature Methods*, 12(10):931–934, 2015. doi: 10.1038/nmeth.3547.

Zhihan Zhou, Yanrong Ji, Weijian Li, Pratik Dutta, Ramana V. Davuluri, and Han Liu. DNABERT-2: Efficient foundation model and benchmark for multi-species genomes. In *International Conference on Learning Representations*, 2024.

Zhihan Zhou, Weimin Wu, Harrison Ho, Jiayi Wang, Lizhen Shi, Ramana V. Davuluri, Zhong Wang, and Han Liu. DNABERT-S: Pioneering species differentiation with species-aware DNA embeddings. *Bioinformatics*, 41(Supplement 1):i255–i264, 2025. doi: 10.1093/bioinformatics/btaf188.

Andrzej Zielezinski, Susana Vinga, Jonas Almeida, and Wojciech M. Karlowski. Alignment-free sequence comparison: Benefits, applications, and tools. *Genome Biology*, 18:186, 2017. doi: 10.1186/s13059-017-1319-7.

A PROOFS

A.1 PROOF OF PROPOSITION 1: COMPLETENESS

For a sequence of fixed length N, let

$$R_\star = \left\lfloor \frac{N}{2} \right\rfloor .$$

For each nucleotide $b \in \mathcal{B}$, Eq. 2 defines the normalized forward discrete Fourier transform

$$F_b(k) = \frac{1}{N} \sum_{t=0}^{N-1} x_b(t) \exp\left(\frac{2\pi i k t}{N} \right). \tag{10}$$

Under this convention, the inverse transform is

$$x_b(t) = \sum_{k=0}^{N-1} F_b(k) \exp\left(-\frac{2\pi i k t}{N} \right). \tag{11}$$

Because $x_b(t)$ is real-valued, its Fourier coefficients satisfy the conjugate-symmetry relation

$$F_b(N-k) = \overline{F_b(k)}. \tag{12}$$

Hence the coefficients indexed by

$$k = 0, \ldots, R_\star$$

contain all nonnegative-frequency coefficients needed to reconstruct the omitted negative frequencies. When N is even, the Nyquist coefficient $F_b(N/2)$ is real-valued and is its own conjugate.

Therefore, the coefficients retained by the full-resolution RNS determine all omitted negative-frequency coefficients through Eq. 12. Substituting the complete spectrum into Eq. 11 uniquely recovers every indicator signal $x_b(t)$.

This representation is conjugacy-reduced but not coordinate-minimal. The Nyquist imaginary coordinate is identically zero for even N, and the one-hot channel identity $\sum_b x_b(t) = 1$ induces linear relations among nucleotide-channel coefficients. These redundancies do not affect invertibility.

For a canonical DNA sequence, exactly one of

$$x_A(t),\ x_C(t),\ x_G(t),\ x_T(t)$$

equals one at each position t, while the remaining three equal zero. Recovering the four indicator signals therefore uniquely identifies the nucleotide at every position. Thus, for fixed N, the full nonnegative-frequency RNS uniquely determines the original sequence.

A.2 PROOF OF PROPOSITION 2: SPECTRAL–HAMMING IDENTITY

Let S and T be two canonical DNA sequences of equal length N. For each nucleotide channel, define

$$\Delta x_b(t) = x_b^S(t) - x_b^T(t), \qquad \Delta F_b(k) = F_b^S(k) - F_b^T(k). \tag{13}$$

Under the normalization in Eq. 2, Parseval's identity gives

$$\sum_{k=0}^{N-1} |\Delta F_b(k)|^2 = \frac{1}{N} \sum_{t=0}^{N-1} |\Delta x_b(t)|^2 . \tag{14}$$

We now sum over the four nucleotide channels. At a position where S and T contain the same nucleotide, all four indicator differences are zero. At a mismatched position, suppose for example that $S_t = A$ and $T_t = C$. Then

$$\Delta x_A(t) = 1, \qquad \Delta x_C(t) = -1,$$

while the other two channel differences are zero. Therefore every mismatch contributes

$$\sum_{b \in \mathcal{B}} |\Delta x_b(t)|^2 = 1^2 + (-1)^2 = 2. \tag{15}$$

Summing over positions yields

$$\sum_{b\in\mathcal{B}}\sum_{t=0}^{N-1}|\Delta x_b(t)|^2 = 2\,d_H(S,T), \tag{16}$$

where $d_H(S,T)$ is the Hamming distance.

Combining Eqs. 14 and 16 gives

$$\sum_{b\in\mathcal{B}}\sum_{k=0}^{N-1}|\Delta F_b(k)|^2 = \frac{2}{N}d_H(S,T). \tag{17}$$

Since $\Delta x_b(t)$ is real-valued, $\Delta F_b(k)$ also satisfies conjugate symmetry. Hence conjugate frequency pairs have equal squared magnitude,

$$|\Delta F_b(N-k)|^2 = |\Delta F_b(k)|^2\,. \tag{18}$$

The complete spectral sum can therefore be written using only the nonnegative-frequency half-spectrum:

$$\sum_{k=0}^{N-1}|\Delta F_b(k)|^2 = \sum_{k=0}^{R_\star} w_k\,|\Delta F_b(k)|^2\,, \tag{19}$$

where $w_0 = 1$, $w_k = 2$ for conjugate frequency pairs, and $w_{N/2} = 1$ for the Nyquist term when N is even.

Using the definition of D_R^2 in Eq. 4, Eqs. 17 and 19 give

$$D_{R_\star}^2(S,T) = \frac{2}{N}d_H(S,T). \tag{20}$$

Finally, for every $R < R_\star$,

$$D_{R+1}^2(S,T) - D_R^2(S,T) = \sum_{b\in\mathcal{B}} w_{R+1}\left|F_b^S(R+1) - F_b^T(R+1)\right|^2 \tag{21}$$

$$\geq 0. \tag{22}$$

Therefore,

$$D_0^2(S,T) \leq D_1^2(S,T) \leq \cdots \leq D_{R_\star}^2(S,T), \tag{23}$$

which proves both the nested-distance property and the exact proportionality to Hamming distance at full resolution.

A.3 Reverse-Complement Transformation

Let $\bar{b}$ denote the complement of nucleotide b, and let S^{RC} denote the reverse complement of S. The corresponding indicator signals satisfy

$$x_b^{S^{\mathrm{RC}}}(t) = x_{\bar{b}}^S(N-1-t). \tag{24}$$

Using Eq. 2,

$$F_b^{S^{\mathrm{RC}}}(k) = \frac{1}{N}\sum_{t=0}^{N-1} x_{\bar{b}}^S(N-1-t)\exp\left(\frac{2\pi ikt}{N}\right). \tag{25}$$

Let

$$u = N-1-t.$$

Then $t = N-1-u$, and reordering the sum gives

$$F_b^{S^{\mathrm{RC}}}(k) = \frac{1}{N}\sum_{u=0}^{N-1} x_{\bar{b}}^S(u)\exp\left(\frac{2\pi ik(N-1-u)}{N}\right) \tag{26}$$

$$= \exp\left(-\frac{2\pi ik}{N}\right)\frac{1}{N}\sum_{u=0}^{N-1} x_{\bar{b}}^S(u)\exp\left(-\frac{2\pi iku}{N}\right). \tag{27}$$

Because $x_{\tilde{b}}^{S}(u)$ is real-valued,

$$\frac{1}{N}\sum_{u=0}^{N-1} x_{\tilde{b}}^{S}(u)\exp\left(-\frac{2\pi iku}{N}\right) = \overline{F_{\tilde{b}}^{S}(k)}. \tag{28}$$

Therefore,

$$F_{b}^{S^{\mathrm{RC}}}(k) = \exp\left(-\frac{2\pi ik}{N}\right)\overline{F_{\tilde{b}}^{S}(k)}, \tag{29}$$

which proves Eq. 7. Thus reverse complementation induces a fixed nucleotide-channel permutation, complex conjugation, and a frequency-dependent phase factor in the RNS representation.

B Implementation details

Tokenizer and positional coordinates. The reverse-complement-closed BPE tokenizer contains 3,991 deployed tokens. Each token retains the base-pair interval from which it was produced. Rotary position embeddings are evaluated at token centers in base-pair coordinates, and sequence-level representations use the span-weighted pooling defined in Eq. 8. Special tokens and padding are excluded from pooling.

RNS target construction. RNS targets are computed from the clean sequence after reverse-complement augmentation and before MLM corruption. When tokenization reaches the 512-token limit, the target is computed from the clean prefix actually covered by the tokenizer. Thus the encoder receives a corrupted tokenized version of the same visible sequence from which the analytical target is calculated.

The implementation computes Fourier coefficients directly using cosine and sine projections,

$$\mathrm{Re}\, F_b(k) = \frac{1}{N}\sum_{t=0}^{N-1} x_b(t)\cos\left(\frac{2\pi kt}{N}\right), \tag{30}$$

$$\mathrm{Im}\, F_b(k) = \frac{1}{N}\sum_{t=0}^{N-1} x_b(t)\sin\left(\frac{2\pi kt}{N}\right). \tag{31}$$

For RNS-4, each nucleotide contributes

$$[p_b, \mathrm{Re}\, F_b(1), \mathrm{Im}\, F_b(1), \ldots, \mathrm{Re}\, F_b(4), \mathrm{Im}\, F_b(4)],$$

giving 36 dimensions in total. Apart from the $1/N$ Fourier normalization, no additional standardization, clipping, or per-dimension scaling is applied to the training targets.

Auxiliary loss (Global RNS). The RNS prediction head maps the 768-dimensional sequence representation to 512 dimensions, followed by GELU and layer normalization, and then projects to the 36-dimensional target. Training uses Smooth L1 loss with the default $\beta = 1$ and mean reduction. The auxiliary weight increases linearly to $\lambda = 50$ during the first 5,000 optimization steps.

C Downstream and Descriptor Protocols

Full finetuning. Each checkpoint is finetuned independently on every task with seeds 41, 42, and 43 for three epochs. The official training split is divided once per seed into a stratified 90% training subset and 10% validation subset. We select the epoch with the highest validation MCC and evaluate the official test split once. All encoder parameters and the dropout–linear classification head are updated. Optimization uses AdamW, batch size 8, learning rate 2×10^{-5}, weight decay 0.01, 10% linear warm-up followed by cosine decay, maximum gradient norm 1.0, and dropout 0.1. The training objective is class-balanced cross-entropy. Token representations are aggregated with Eq. 8; encoder inputs are limited to 512 tokens.

Table 2: Main pretraining configuration.

Item	Value
Layers / hidden / heads	12 / 768 / 12
Feed-forward width	3,072 (GEGLU)
BPE vocabulary	3,991 deployed tokens
Window / token limit	1,000 bp / 512 tokens
Deployment parameters	116,352,770
MLM probability	0.15
RC augmentation probability	0.5
Training steps	150,000
Pretraining seeds	42, 43, 44
Effective batch	1,024 sequences
Optimizer	AdamW
Peak learning rate	2×10^{-4}
Warm-up / schedule	2,000 steps / cosine decay
Weight decay	0.01
RNS order / dimension	4 / 36
RNS weight / warm-up	50 / 5,000 steps
Precision / devices	bfloat16 / 2× NVIDIA H200
Prepared pretraining windows	35,605,066
Prepared canonical bases	35,605,049,541

Fixed-descriptor evaluation. The RNS resolution and component sweeps use the official train and test splits of all 18 tasks. For each of split seeds 41, 42, and 43, a stratified 90/10 split of the official training set selects the inverse regularization strength $C \in \{0.01, 0.1, 1, 10\}$ by validation MCC. The selected pipeline is then refitted on the complete official training split and evaluated once on the official test split. The classifier is ℓ_2-regularized logistic regression with class-balanced weights and maximum 3,000 iterations. A `StandardScaler(with_mean=False)` is fitted inside the training pipeline. No sequence subsampling is used. RNS-4 is the compact target used in all reported SpanDNA pretraining runs. The sweep characterizes how predictive information changes with spectral resolution; it is not used to select a pretraining optimum.

D Representation-Probing Protocol

The analysis uses the seed-42 MLM-only and RNS-supervised pretraining pair from Table 1b.

Layer-wise probing uses the H3K4me1 benchmark. From the official training split, 5,000 sequences are sampled without replacement using seed 42; from the official test split, 2,000 sequences are sampled without replacement using seed 43. The test sequences are not used for probe fitting or feature standardization.

The pretrained encoder is frozen and evaluated without MLM corruption or additional reverse-complement augmentation. At each encoder layer, token representations are aggregated using the same base-pair-span-weighted pooling as Eq. 8. A `StandardScaler` is fitted only to the 5,000 training representations, followed by ridge regression with fixed coefficient $\alpha = 1$. No validation set or cross-validation is used to select α.

The probe jointly predicts all 36 RNS dimensions. Reported spectral probing results use the 32 non-DC real and imaginary coordinates, excluding the four nucleotide-composition dimensions. R^2 is computed using variance-weighted multi-output aggregation. The final-layer non-DC R^2 values are -0.1760 for MLM-only pretraining and 0.9089 for RNS-supervised pretraining. For the four DC composition coordinates, the corresponding final-layer values are 0.9869 and 0.9936.

For auxiliary-head diagnostics, the 2,000 official test sequences yield

$$2{,}000 \times 32 = 64{,}000$$

non-DC scalar coefficient predictions. Similarity-map visualizations use a deterministic subset of 256 of these sequences. Similarities are computed from the 32 non-DC coordinates after dimension-wise standardization using statistics from the 2,000 analytical targets and row-wise ℓ_2 normalization.

E STATISTICAL REPORTING

For each pretraining seed (42, 43, or 44), task, and objective, the three finetuning seeds are first averaged. Mean and sample standard deviation are then computed across the three pretraining seeds from these seed-level task averages.

For task-level paired analyses, each task is first averaged over all three pretraining seeds and finetuning seeds. The resulting 18 paired task differences are used to compute the mean, median, win/tie/loss count, and two-sided Wilcoxon signed-rank test. The 95% task-bootstrap interval is obtained from 10,000 paired-task bootstrap resamples using random seed 2026. This interval conditions on the observed seed-averaged task effects. It summarizes variation across the evaluated task set rather than jointly propagating pretraining and finetuning uncertainty; variability across pretraining repetitions is reported separately in Table 1b.

F ADDITIONAL DOWNSTREAM RESULTS

F.1 DETAILED COMPARISON WITH EXISTING MODELS

Tables 3–5 report the full per-task MCC values underlying the benchmark comparison in Section 4.4. For locally rerun baselines, entries are mean ± SD over three finetuning seeds under the three-epoch protocol. For SpanDNA, each task is first averaged over the three finetuning seeds within each pretrained checkpoint, and entries report mean ± sample SD across pretraining seeds 42, 43, and 44, consistent with Table 1a and Table 6. Published baselines are taken from PatchDNA Supplementary Table 11 and retain the reported three-seed values.

Table 3: Per-task test MCC comparison (I/III). For SpanDNA, values are mean ± SD across three pretraining seeds after averaging three finetuning seeds per checkpoint; other locally rerun models report mean ± SD over three finetuning seeds, and published baselines retain their reported values.

Model	Promoter all	Promoter no-TATA	Promoter TATA	Enhancer	Enhancer type	Splice accept.
SpanDNA	0.742 ± 0.007	0.769 ± 0.007	0.786 ± 0.009	0.535 ± 0.002	0.488 ± 0.002	0.886 ± 0.004
DNABERT-2 (117M)	0.760 ± 0.005	0.771 ± 0.007	0.829 ± 0.015	0.539 ± 0.007	0.497 ± 0.007	0.842 ± 0.008
ModernGENA-Base	0.758 ± 0.008	0.768 ± 0.006	0.859 ± 0.009	0.528 ± 0.028	0.490 ± 0.017	0.824 ± 0.010
MxDNA	0.734 ± 0.013	0.755 ± 0.010	0.831 ± 0.038	0.519 ± 0.014	0.480 ± 0.010	0.812 ± 0.032
PatchDNA	0.791 ± 0.009	0.788 ± 0.005	0.840 ± 0.019	0.528 ± 0.009	0.496 ± 0.008	0.754 ± 0.040
DNABERT-1 (6-mer)	0.736 ± 0.010	0.741 ± 0.005	0.840 ± 0.025	0.532 ± 0.010	0.491 ± 0.009	0.954 ± 0.002
PatchDNA Entropy	0.725 ± 0.008	0.730 ± 0.007	0.785 ± 0.016	0.523 ± 0.001	0.488 ± 0.008	0.868 ± 0.015
Caduceus-ps	0.742 ± 0.010	0.764 ± 0.013	0.761 ± 0.028	0.510 ± 0.017	0.471 ± 0.006	0.765 ± 0.006
GENA-LM-Base	0.738 ± 0.007	0.736 ± 0.025	0.689 ± 0.038	0.483 ± 0.023	0.467 ± 0.012	0.760 ± 0.005
HyenaDNA	0.693 ± 0.007	0.724 ± 0.004	0.831 ± 0.057	0.479 ± 0.005	0.450 ± 0.003	0.820 ± 0.015
PhyloP	0.405 ± 0.002	0.393 ± 0.006	0.469 ± 0.006	0.181 ± 0.007	0.167 ± 0.002	0.543 ± 0.001

Table 4: Per-task test MCC comparison (II/III). For SpanDNA, values are mean ± SD across three pretraining seeds after averaging three finetuning seeds per checkpoint; other locally rerun models report mean ± SD over three finetuning seeds, and published baselines retain their reported values.

Model	Splice all	Splice donor	H2AFZ	H3K27ac	H3K27me3	H3K36me3
SpanDNA	0.879 ± 0.001	0.931 ± 0.002	0.513 ± 0.004	0.510 ± 0.006	0.620 ± 0.001	0.638 ± 0.003
DNABERT-2 (117M)	0.848 ± 0.007	0.848 ± 0.007	0.515 ± 0.002	0.513 ± 0.018	0.612 ± 0.005	0.646 ± 0.006
ModernGENA-Base	0.837 ± 0.007	0.846 ± 0.006	0.506 ± 0.005	0.497 ± 0.018	0.606 ± 0.013	0.619 ± 0.015
MxDNA	0.860 ± 0.007	0.931 ± 0.021	0.512 ± 0.003	0.489 ± 0.031	0.599 ± 0.015	0.618 ± 0.002
PatchDNA	0.760 ± 0.019	0.706 ± 0.026	0.523 ± 0.010	0.486 ± 0.015	0.607 ± 0.008	0.621 ± 0.007
DNABERT-1 (6-mer)	0.946 ± 0.014	0.962 ± 0.005	0.434 ± 0.011	0.441 ± 0.009	0.562 ± 0.004	0.519 ± 0.010
PatchDNA Entropy	0.884 ± 0.013	0.654 ± 0.016	0.521 ± 0.009	0.484 ± 0.035	0.595 ± 0.004	0.584 ± 0.020
Caduceus-ps	0.796 ± 0.021	0.771 ± 0.013	0.507 ± 0.007	0.475 ± 0.021	0.591 ± 0.009	0.607 ± 0.008
GENA-LM-Base	0.764 ± 0.013	0.781 ± 0.004	0.466 ± 0.035	0.495 ± 0.010	0.588 ± 0.004	0.602 ± 0.021
HyenaDNA	0.849 ± 0.006	0.840 ± 0.029	0.481 ± 0.005	0.440 ± 0.003	0.554 ± 0.014	0.549 ± 0.002
PhyloP	0.283 ± 0.004	0.547 ± 0.001	−0.017 ± 0.062	0.105 ± 0.028	0.233 ± 0.032	0.304 ± 0.003

Table 5: Per-task test MCC comparison (III/III). For SpanDNA, values are mean ± SD across three pretraining seeds after averaging three finetuning seeds per checkpoint; other locally rerun models report mean ± SD over three finetuning seeds, and published baselines retain their reported values.

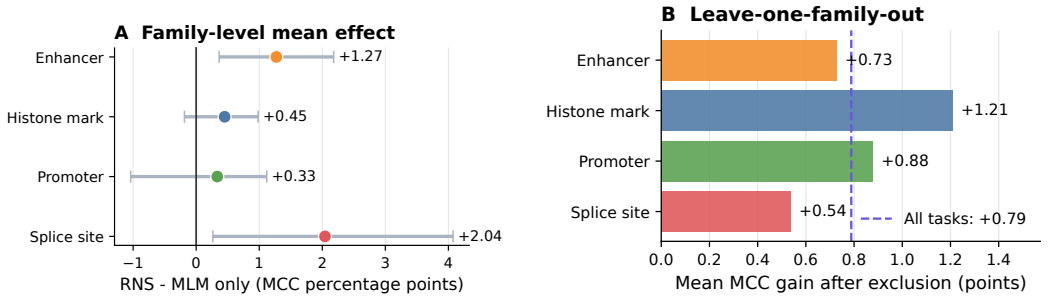
Model	H3K4me1	H3K4me2	H3K4me3	H3K9ac	H3K9me3	H4K20me1	Mean
SpanDNA	0.501 ± 0.002	0.566 ± 0.001	0.625 ± 0.001	0.572 ± 0.006	0.478 ± 0.004	0.661 ± 0.004	0.6500
DNABERT-2 (117M)	0.511 ± 0.010	0.592 ± 0.009	0.636 ± 0.010	0.575 ± 0.005	0.488 ± 0.035	0.673 ± 0.002	0.6496
ModernGENA-Base	0.478 ± 0.010	0.560 ± 0.011	0.645 ± 0.004	0.561 ± 0.013	0.461 ± 0.017	0.662 ± 0.008	0.6392
MxDNA	0.497 ± 0.001	0.563 ± 0.012	0.627 ± 0.017	0.534 ± 0.015	0.467 ± 0.023	0.646 ± 0.007	0.6374
PatchDNA	0.480 ± 0.003	0.573 ± 0.004	0.634 ± 0.005	0.569 ± 0.010	0.470 ± 0.017	0.637 ± 0.007	0.6257
DNABERT-1 (6-mer)	0.422 ± 0.010	0.497 ± 0.013	0.619 ± 0.003	0.520 ± 0.008	0.421 ± 0.004	0.574 ± 0.016	0.6228
PatchDNA Entropy	0.472 ± 0.011	0.568 ± 0.021	0.589 ± 0.010	0.546 ± 0.009	0.473 ± 0.019	0.626 ± 0.027	0.6175
Caduceus-ps	0.471 ± 0.014	0.565 ± 0.008	0.617 ± 0.009	0.526 ± 0.009	0.435 ± 0.015	0.639 ± 0.009	0.6118
GENA-LM-Base	0.465 ± 0.014	0.538 ± 0.027	0.610 ± 0.055	0.525 ± 0.007	0.440 ± 0.009	0.644 ± 0.011	0.5995
HyenaDNA	0.438 ± 0.007	0.523 ± 0.025	0.618 ± 0.007	0.497 ± 0.014	0.371 ± 0.026	0.617 ± 0.008	0.5986
PhyloP	0.006 ± 0.041	-0.020 ± 0.058	0.009 ± 0.084	0.026 ± 0.041	0.072 ± 0.035	0.059 ± 0.103	0.2092

F.2 Complete metrics

Table 6: Complete downstream metrics. Within each task and pretraining seed, three finetuning seeds are averaged first; entries report mean ± sample SD across pretraining seeds 42, 43, and 44. AUROC is binary AUROC for binary tasks and macro one-vs-rest AUROC for multiclass tasks.

	MCC		Macro F1		Accuracy		AUROC	
Task	MLM	MLM + RNS	MLM	MLM + RNS	MLM	MLM + RNS	MLM	MLM + RNS
enhancers	0.5137 ± 0.0046	0.5355 ± 0.0016	0.7546 ± 0.0038	0.7642 ± 0.0005	0.7548 ± 0.0034	0.7646 ± 0.0006	0.8331 ± 0.0015	0.8401 ± 0.0012
enhancer types	0.4843 ± 0.0019	0.4880 ± 0.0021	0.5633 ± 0.0057	0.5610 ± 0.0019	0.7210 ± 0.0016	0.7227 ± 0.0012	0.7940 ± 0.0020	0.7940 ± 0.0001
H2AFZ	0.5150 ± 0.0019	0.5133 ± 0.0036	0.7510 ± 0.0016	0.7461 ± 0.0088	0.7529 ± 0.0011	0.7495 ± 0.0063	0.8344 ± 0.0004	0.8425 ± 0.0010
H3K27ac	0.5015 ± 0.0037	0.5102 ± 0.0055	0.7499 ± 0.0012	0.7542 ± 0.0025	0.7502 ± 0.0014	0.7545 ± 0.0026	0.8369 ± 0.0016	0.8382 ± 0.0022
H3K27me3	0.6183 ± 0.0030	0.6198 ± 0.0012	0.7978 ± 0.0014	0.7990 ± 0.0008	0.8003 ± 0.0014	0.8015 ± 0.0007	0.8813 ± 0.0006	0.8795 ± 0.0002
H3K36me3	0.6242 ± 0.0009	0.6377 ± 0.0030	0.8084 ± 0.0015	0.8145 ± 0.0022	0.8093 ± 0.0012	0.8156 ± 0.0020	0.8911 ± 0.0011	0.8935 ± 0.0003
H3K4me1	0.4954 ± 0.0013	0.5007 ± 0.0019	0.7412 ± 0.0004	0.7411 ± 0.0043	0.7436 ± 0.0005	0.7444 ± 0.0030	0.8233 ± 0.0011	0.8278 ± 0.0025
H3K4me2	0.5619 ± 0.0072	0.5656 ± 0.0009	0.7764 ± 0.0044	0.7784 ± 0.0009	0.7777 ± 0.0041	0.7796 ± 0.0007	0.8655 ± 0.0004	0.8638 ± 0.0014
H3K4me3	0.6221 ± 0.0060	0.6252 ± 0.0014	0.8106 ± 0.0031	0.8118 ± 0.0011	0.8107 ± 0.0030	0.8120 ± 0.0010	0.9001 ± 0.0004	0.9004 ± 0.0017
H3K9ac	0.5607 ± 0.0031	0.5724 ± 0.0062	0.7791 ± 0.0019	0.7851 ± 0.0032	0.7794 ± 0.0018	0.7854 ± 0.0032	0.8575 ± 0.0016	0.8687 ± 0.0001
H3K9me3	0.4961 ± 0.0070	0.4775 ± 0.0042	0.7454 ± 0.0038	0.7364 ± 0.0012	0.7464 ± 0.0035	0.7373 ± 0.0014	0.8260 ± 0.0039	0.8218 ± 0.0015
H4K20me1	0.6429 ± 0.0104	0.6606 ± 0.0036	0.8142 ± 0.0063	0.8254 ± 0.0016	0.8158 ± 0.0060	0.8264 ± 0.0016	0.8939 ± 0.0046	0.9024 ± 0.0018
promoter: all	0.7307 ± 0.0014	0.7419 ± 0.0066	0.8651 ± 0.0006	0.8706 ± 0.0031	0.8652 ± 0.0006	0.8707 ± 0.0032	0.9419 ± 0.0009	0.9453 ± 0.0030
promoter: no TATA	0.7594 ± 0.0024	0.7686 ± 0.0065	0.8794 ± 0.0012	0.8842 ± 0.0033	0.8794 ± 0.0012	0.8842 ± 0.0033	0.9532 ± 0.0006	0.9545 ± 0.0007
promoter: TATA	0.7968 ± 0.0048	0.7865 ± 0.0093	0.8983 ± 0.0024	0.8925 ± 0.0048	0.8983 ± 0.0024	0.8926 ± 0.0048	0.9570 ± 0.0032	0.9558 ± 0.0025
splice: acceptors	0.8457 ± 0.0046	0.8865 ± 0.0041	0.9221 ± 0.0023	0.9430 ± 0.0021	0.9221 ± 0.0023	0.9430 ± 0.0021	0.9736 ± 0.0009	0.9855 ± 0.0003
splice: all	0.8611 ± 0.0082	0.8789 ± 0.0013	0.9067 ± 0.0053	0.9192 ± 0.0009	0.9071 ± 0.0053	0.9191 ± 0.0009	0.9806 ± 0.0022	0.9843 ± 0.0004
splice: donors	0.9280 ± 0.0005	0.9306 ± 0.0018	0.9638 ± 0.0002	0.9652 ± 0.0009	0.9638 ± 0.0002	0.9652 ± 0.0009	0.9904 ± 0.0003	0.9926 ± 0.0004

F.3 Family robustness



(a) Family effects.

(b) Leave-one-family-out.

Figure 5: Task-family diagnostics after averaging pretraining seeds 42, 43, and 44 and the three finetuning seeds. Intervals in (a) resample the aggregated task effects within each family and are necessarily wide for families containing only two or three tasks. All leave-one-family-out mean effects remain positive.

Table 7: Family-level RNS minus MLM MCC differences after averaging pretraining seeds 42, 43, and 44 and the three finetuning seeds. Values are percentage points; intervals are 95% task-bootstrap intervals.

Family	Tasks	Mean	95% CI	W/T/L
Enhancer	2	+1.27	$[+0.36, +2.18]$	2/0/0
Histone mark	10	+0.45	$[-0.19, +0.99]$	8/0/2
Promoter	3	+0.33	$[-1.04, +1.12]$	2/0/1
Splice site	3	+2.04	$[+0.27, +4.07]$	3/0/0

F.4 PER-TASK RESULTS

Table 8: Aggregate per-task RNS minus MLM MCC differences after averaging pretraining seeds 42, 43, and 44 and the three finetuning seeds. Values are percentage points.

Task	RNS − MLM
splice_sites_acceptors	+4.07
enhancers	+2.18
splice_sites_all	+1.79
H4K20me1	+1.78
H3K36me3	+1.35
H3K9ac	+1.17
promoter_all	+1.12
promoter_no_tata	+0.92
H3K27ac	+0.88
H3K4me1	+0.53
H3K4me2	+0.37
enhancers_types	+0.36
H3K4me3	+0.31
splice_sites_donors	+0.27
H3K27me3	+0.15
H2AFZ	−0.16
promoter_tata	−1.04
H3K9me3	−1.85

G ADDITIONAL TRAINING DIAGNOSTICS AND INFERENCE MEASUREMENTS

G.1 AUXILIARY-TARGET TRAINING DYNAMICS

Target-comparison protocol. The composition, Global RNS, and multiscale RNS runs use the same architecture, data configuration, optimization budget, and downstream evaluation protocol. The targets differ in dimensionality, so auxiliary-loss values are shown only to document convergence; downstream MCC provides the common evaluation measure across target variants.

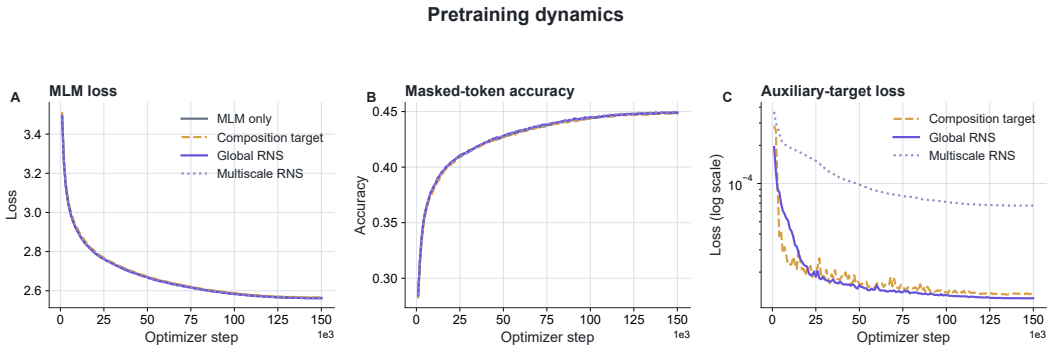


Figure 6: MLM convergence and auxiliary losses for the retained pretraining objectives. MLM loss and masked-token accuracy remain closely matched. Auxiliary-loss magnitudes are not directly comparable because target dimensions differ.

G.2 INFERENCE MEASUREMENTS

The RNS prediction head is used only during pretraining and is discarded afterward. Consequently, MLM-only and RNS-supervised checkpoints share the same encoder architecture and parameter count during downstream evaluation. Encoder latency and peak allocated GPU memory are therefore measured using the same active model architecture for both variants, with no additional RNS-specific computation introduced during downstream finetuning or inference.

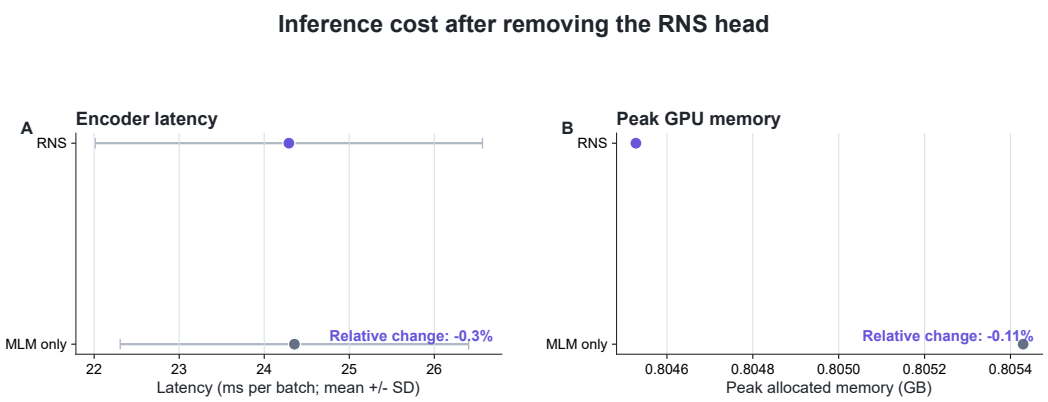


Figure 7: Matched encoder inference measurements after removing the RNS head. Observed latency and memory differences are within measurement noise.

H MULTISCALE RNS TARGET

The multiscale target used in Table 1c concatenates RNS-4 computed on the full sequence, its two approximately equal contiguous partitions, and its four approximately equal contiguous partitions. Each segment contributes a 36-dimensional RNS representation, giving

$$1 \times 36 + 2 \times 36 + 4 \times 36 = 252$$

dimensions. Within each scale, segments are ordered from left to right. This construction changes spatial partitioning while retaining $R = 4$ within each segment; it is not a sweep of the global spectral order.