EMERGENT CODON FRAME REPRESENTATIONS REVEAL MECHANISMS OF TRANSLATION IN GENOMIC FOUNDATION MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Interpreting scientific foundation models requires both identifying scientific variables in model representations and showing that the model uses those variables in its computation. As a test case with known ground truth, we first investigate how the triplet structure of coding RNA emerges in the representations of four genomic foundation models. Across all four models, representations of RNA coding regions exhibit stronger period-3 structure than non-coding regions, and in three of the four, coding region representations projected onto their top two principal components reorganize from nucleotide identity toward codon phase with depth. We use these findings as motivation to look more closely at the RNA-to-protein translation mechanism in MIMIC, a multimodal encoder-decoder foundation model trained jointly on RNA and protein modalities, and show that the codon phase representation is causally used for protein decoding. Rotating RNA representations from MIMIC's encoder by 120° in a two-dimensional codon phase subspace shifts the model's reading frame for protein decoding by one nucleotide. The inverse rotation induces the opposite shift and repeated rotations compose according to the expected modulo-three structure. Tracing the computation downstream, we find that amino acid identity becomes accessible through the decoder's own cross-attention channel late in encoder representations. A single cross-attention head in the decoder's final block reads the amino acid from the encoder's output representation at the codon's middle nucleotide, which it preferentially attends to, and supplying this position alone recovers the correct amino acid in the single-residue-masked assay. In a natural transcript with overlapping protein-coding reading frames, phase rotation and start-site conditioning separately control translation register and origin, together recovering the alternative protein. These results characterize mechanisms of protein translation in MIMIC and link an emergent biological coordinate to the computations that determine protein output.

1 INTRODUCTION AND RELATED WORK

Biological foundation models increasingly support diverse predictive and generative tasks, and recent interpretability work has begun to recover meaningful biological structure from their internal representations. For example, interpretability analyses have identified features associated with protein structure, splice boundaries, and transcription factor motifs (Adams et al., 2025; Lu et al., 2026; Candido et al., 2026; Brixi et al., 2026). Interpreting more complex learned behavior remains challenging however because an apparently meaningful feature can be difficult to attribute or causally validate. Rather than beginning with an unknown biological mechanism, we instead start from a well-characterized process and ask whether the model reconstructs and uses it.

The process of translating from RNA to protein is a natural test case. During translation, ribosomes read coding messenger RNA in non-overlapping triplets, called codons, which specify amino acids or translation termination (Crick et al., 1961) (see Fig. 1). Translated coding regions therefore obey a three-base organization that is absent from most non-coding sequence, and the structure of the genetic code makes the three positions within each codon statistically distinct. These asymmetries give rise to characteristic position-specific nucleotide frequencies and a period-3 sequence signature, which have been exploited to identify coding sequence through nucleotide-frequency statistics

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.

(Fickett, 1982) and Fourier analysis (Tiwari et al., 1997). Reading frame specifies which of three nucleotide offsets defines the triplets used for translation, and we refer to a nucleotide's position within this register (or its modulo-3 position with respect to the start of the spliced coding region) as its codon phase.

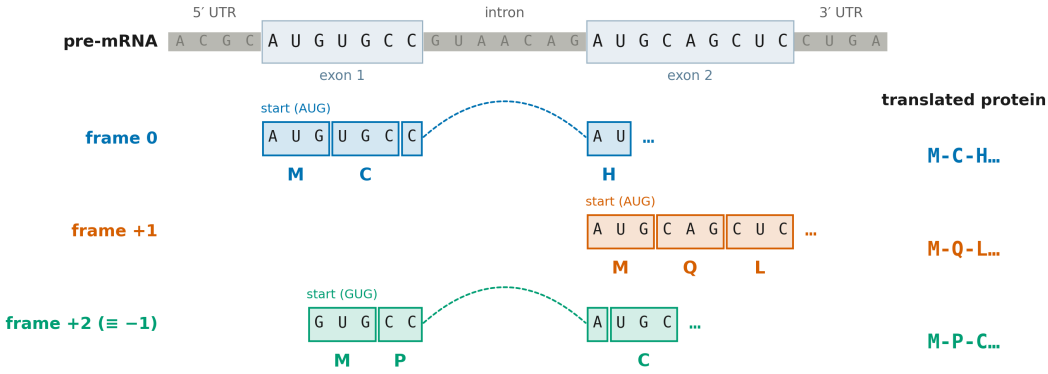

Figure 1: **Reading frame determines codon grouping.** Exons are retained and the intron is removed during splicing. Untranslated regions (UTRs) flank the coding region. The illustrated start sites select different triplet registers in the same spliced RNA. A codon can span an exon junction.

Structured cyclic variables such as codon phase have appeared in several mechanistic interpretability settings. Neural networks learn group structure in synthetic composition tasks (Chughtai et al., 2023), circular Fourier representations in modular arithmetic (Nanda et al., 2023), and causally relevant circular features for cyclic concepts in language models (Engels et al., 2025). More closely related to reading frame, contextual counting tasks show that Transformers can construct internal coordinates whose meaning depends on locating and tracking a region of context (Golkar et al., 2024), while fixed-width line breaking has been traced to low-dimensional position manifolds that are geometrically transformed to drive model outputs (Gurnee et al., 2025).

Recent genomic foundation models reveal similar behavior, including three-base periodicity in Evo 2 mutational likelihoods (Brixi et al., 2026), codon-periodic chunking in dnaHNet (Shah et al., 2026), and codon phase organization in Carbon embeddings (Ben Allal et al., 2026). LucaVirus further shows that codon phase and amino acid identity are recoverable from nucleotide embeddings, with phase-consistent representations across programmed frameshift boundaries (Pan et al., 2025). These observations establish that genomic models encode codon-scale structure, but leave open how such structure participates in downstream tasks.

In this study, we identify enriched period-three structure across four genomic foundation models and a shift toward codon phase geometry with depth in three of them. In MIMIC (Golkar et al., 2026), we identify a causal pathway from this representation to protein output. Latent rotations induce reading frame offsets, and we isolate a decoder cross-attention head that reads amino acid information from late encoder representations at the middle nucleotide of each codon. We further validate these findings using composed conditioning and rotations to induce alternate protein translation on a natural sequence with multiple possible reading frames.

2 Experimental Setup

Models We analyze four pretrained genomic foundation models spanning different architectures, training objectives, and model scales: MIMIC (Golkar et al., 2026), Evo 2 (Brixi et al., 2026), Nucleotide Transformer v3 (NTv3) (Boshar et al., 2025), and HyenaDNA (Nguyen et al., 2023). Table 1 summarizes the checkpoints and model details used throughout this study. All model parameters are frozen and hidden representations are extracted without fine-tuning.

MIMIC is trained jointly between RNA and protein modalities and can natively translate RNA sequence to amino acid sequence, while the other models tested process nucleic acid data exclusively.

Model	Architecture	Training objective	Params.	RNA→protein
MIMIC	Transformer enc-dec	Multimodal masked modeling	1.0B	Yes
Evo 2	StripedHyena 2	Autoregressive LM	1.11B	No
NTv3	U-Net Transformer	Masked LM	652M	No
HyenaDNA	Hyena	Autoregressive LM	14.2M	No

Table 1: Models analyzed in this study. Parameter counts correspond to the specific checkpoints used in our experiments: `MIMIC`, `evo2_1b_base`, `NTv3_650M_pre`, and `hyenadna-medium-160k-seqlen-hf`.

We therefore use MIMIC for the causal steering and downstream mechanism experiments, where perturbations to RNA representations can be evaluated directly through the model's protein output.

Sequence Data We analyze transcripts spanning diverse microbial and metazoan species and multiple RNA biotypes, including protein-coding and non-coding RNAs. The same sequences are used across all four models to enable direct cross-model comparison. Coding and non-coding regions are defined from reference transcript annotations. Additional spliced metazoan mRNAs are used for MIMIC's RNA-to-protein translation experiments. The transcript cohorts are described in Appendix A.1 and statistical conventions are given in Appendix A.8.

3 Period-3 Structure Emerges in Representations of Coding Regions

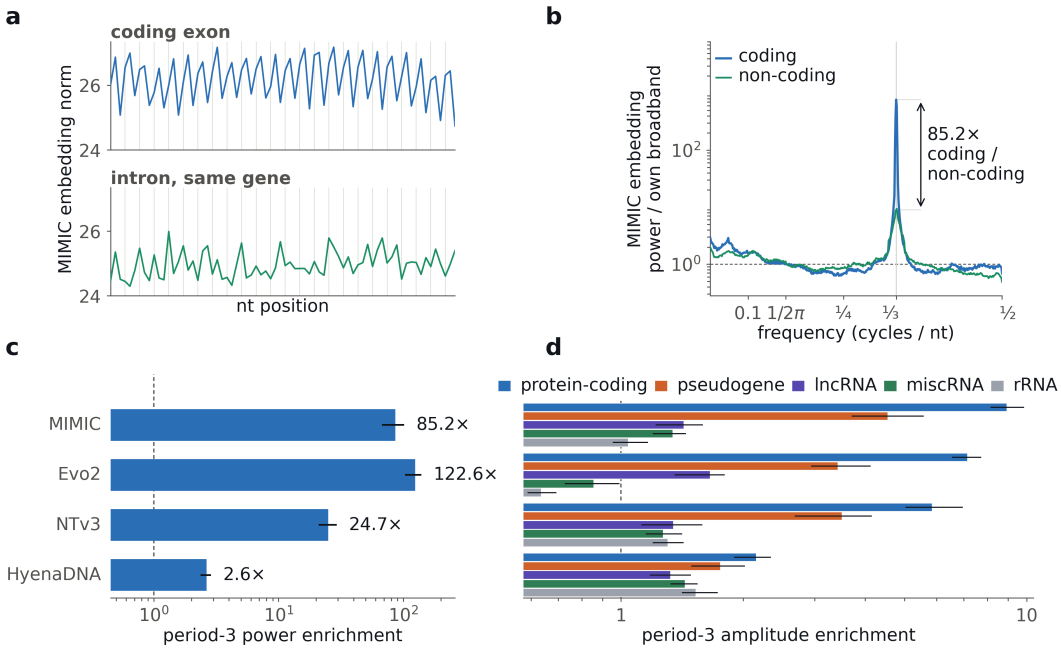

Figure 2: **Period-3 structure is enriched in coding sequence across genomic foundation models.** **a)** MIMIC embedding norm of an exon and intron from one *C. elegans* gene. Vertical rules mark codon boundaries. The norm peaks one nucleotide into each codon across all 24 codons of the exon. **b)** Power spectra of the same quantity over 297 microbial mRNAs and 270 unspliced animal pre-mRNAs. Coding sequence carries $85.2\times$ the period-3 power of non-coding sequence at $f = 1/3$. **c)** Comparing period-3 coding/non-coding power across four genomic foundation models. Each model is read at the layer where its coding versus non-coding contrast is largest (see Fig. A.1). **d)** Period-3 amplitude across metazoan RNA classes.

We characterize sequence representations with two different approaches. First, we measure three-nucleotide periodicity by treating the norm of the hidden state along a sequence as a one-dimensional signal and analyzing its Fourier spectrum. We present results comparing period-3 signal enrichment in coding regions vs. non-coding regions across all four models, and across different functional classes of RNA. Full calculation details are provided in Appendix A.2.

We first ask whether any periodicity is present in a genomic foundation model's latent representations, and whether it is specific to coding sequence. Figure 2a shows the norm $||h||$ of typical coding exon and non-coding intron representations from the output of MIMIC's encoder. We observe a very clear period-3 structure in this coding sample that is not present in the non-coding RNA region. Fig. 2b shows that this is a general effect across our dataset, with an $85.2\times$ stronger peak at 1/3 cycles per nucleotide in coding regions vs. non-coding. This phenomenon follows the triplet codon structure of coding RNA, as shown schematically in Fig. 1.

This enrichment is observed across all four tested genomic foundation models to varying degrees (Fig. 2c). While all models have diverse training data, architectures and parameter counts that could impact this quality, HyenaDNA alone has no attention mechanism, which may lead to the decreased ratio we see here. All models share RNA type character, as we observe the strongest period-3 enrichment among protein coding genes containing annotated coding exons, and the lowest in non-coding RNA classes (Fig. 2d). Interestingly, we observe that pseudogenes, which retain sequence features inherited from protein-coding genes despite generally lacking functional protein-coding capacity, show intermediate period-3 power.

4 Representations Reorient Towards Codon Phase with Depth

Second, we study how nucleotide identity and codon phase are represented across model depth. We use unsupervised PCA to characterize the dominant low-dimensional geometry of the hidden states and quantify its association with each variable using η^2. Separately, linear probes measure whether the same variables are decodable from the higher-dimensional representation. η^2 measures what organizes the leading geometry, whereas probe accuracy measures what information is linearly recoverable. Full fitting and evaluation details are given in Appendix A.3. For each RNA sample, we define codon phase $\phi_k = (k - k_{CDS}) \bmod 3$, where k indexes the spliced mRNA and k_{CDS} is the position of the first base of the coding sequence (CDS) (Fig. 3a).

At MIMIC's embedding layer, unsupervised PCA representations of coding RNA regions are largely organized by nucleotide identity, but as the representations evolve along the encoder blocks, the leading geometry of the coding regions reorganizes to a codon phase-dominant representation (Fig. 3b). Variance in the leading plane explained by nucleotide identity η^2_{nt} decreases in CDS regions and η^2_{phase} increases, reaching 0.84 near the end of the encoder (Fig. 3c), and codon phase is increasingly decodable by supervised linear probes throughout MIMIC's encoder regardless of fluctuations in η^2_{phase}.

For non-coding tokens, η^2_{phase} is small throughout MIMIC's encoder, but the extrapolated codon phase label is decodable above chance in the final layers (Fig. 3d, e) despite not corresponding to a meaningful biological variable in non-coding sequence. Nucleotide identity is decodable consistently for all tokens throughout MIMIC's encoder even though η^2_{nt} decreases near the output layer (Fig. 3f, g). Evo 2 and NTv3 show similar qualitative behavior, as codon phase is clearly separated in representations of coding tokens mid-late in their model depth (see Appendix B.1). HyenaDNA does not show any strong clustering of codon phase, but does separate nucleotide identity strongly across layers.

Because codon phase advances cyclically $\varphi_0 \to \varphi_1 \to \varphi_2 \to \varphi_0$ along a coding sequence, codon phase naturally defines a three-state cyclic variable, and unlike a feature intrinsic to token identity, reading frame is contextual. The same nucleotide can belong to different reading frames, and therefore codons, depending on where translation begins and how the transcript is processed. This makes reading frame a useful intermediate variable for tracing the computation of translation, as we can ask not only whether phase emerges in hidden representations, but whether manipulating that representation produces the corresponding change in protein output.

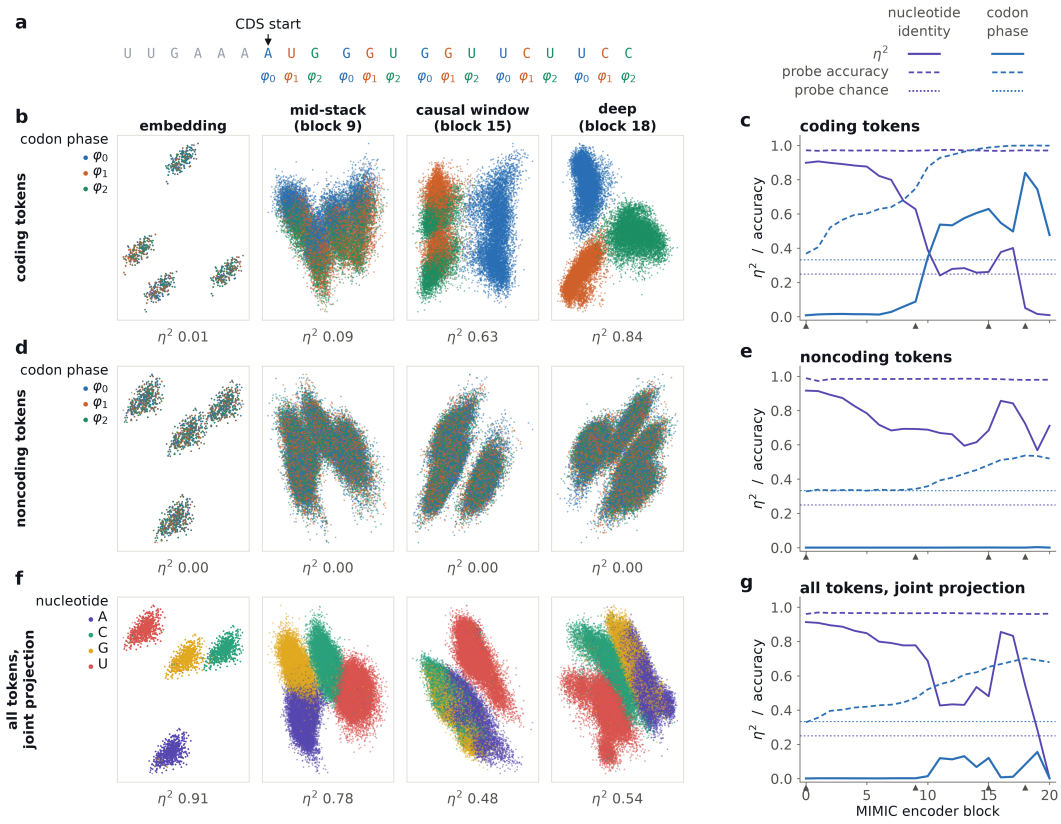

Figure 3: **MIMIC coding sequence representations reorganize from nucleotide identity to codon phase.** **a)** Codon phase φ defined relative to the start of the coding sequence. **b, d, f)** PCA embedding of MIMIC hidden representations across the model's depth. Coding tokens colored by φ in **b**, non-coding tokens colored as if the phase continued out of the coding region in **d**, and all tokens together colored by nucleotide identity in **f**. **c, e, g)** Explained variance η^2 in the PCA plane across encoder blocks in MIMIC, with linear decodability from a 150-principal-component representation shown as a dashed line of the corresponding color. Probe chance is $1/3$ for codon phase and $1/4$ for nucleotide, marked by the dotted lines.

5 Rotating Latent Representations to Causally Steer Reading Frame

To test whether the encoded codon phase information is causally used, we first define a supervised codon phase plane from the three phase centroids on held-out tokens (Fig. 4a). We rotate the component of MIMIC's hidden representations in this plane after one encoder block at a time and observe the change in translation output. Detailed intervention and scoring methods are provided in Appendix A.4.

Fig. 4c shows the strongest steering after encoder blocks 15–17, with a ramp in the preceding blocks and a sharp decline after block 18. This figure also shows that the unsupervised PCA plane used in Section 4 is not effective for causal steering.

As there are only three codon phases, we expect rotations of $\pm 120°$ to be most effective. In fact, only near the midpoints between ideal rotations does the intervention fail to produce a confident translated reading frame (Fig. 4d). At angles other than $\approx \pm 70°$, a rotation snaps onto one of three coding phases corresponding to a discrete reading frame. This effect is shown schematically in Fig. 4b, where rotations of $R = 120°$ induce a forward shift of 1 nucleotide in translation reading frame, and rotations of $R^{-1} = -120°$ induce a reverse shift of -1 nucleotide in translation reading frame. Both

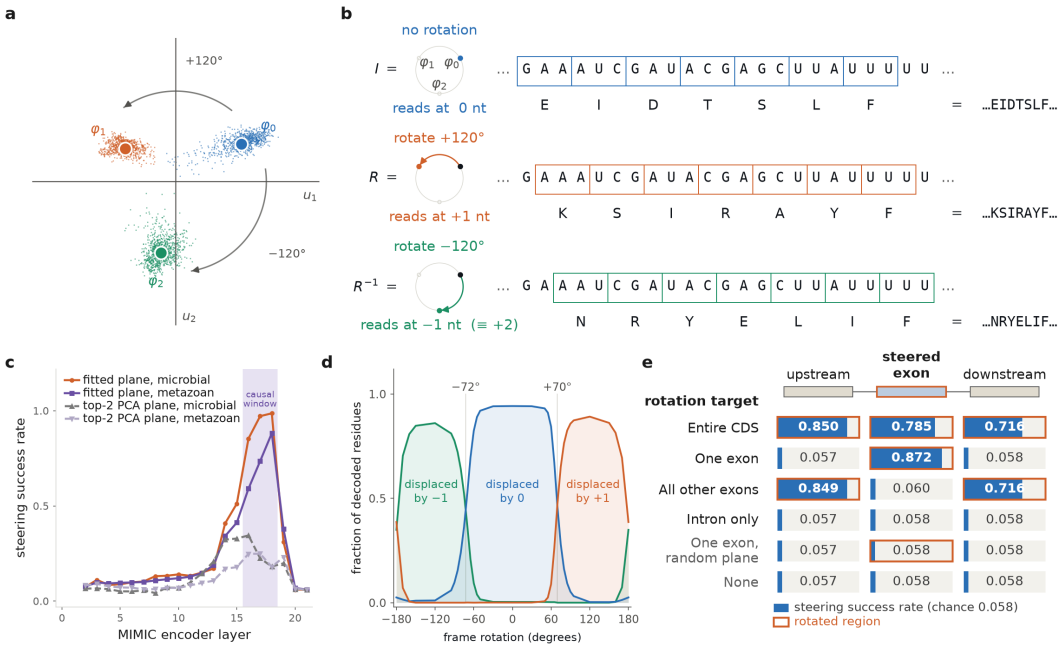

Figure 4: **Codon phase rotations steer MIMIC protein decoding. a)** The three phase centroids in the supervised plane, with held-out coding tokens drawn. **b)** Causal intervention schematic. Rotating MIMIC's hidden representation $\pm 120°$ in the learned codon phase plane induces a ± 1 nt shift in the reading frame used for protein decoding. **c)** Steering success rate across MIMIC encoder layers peaking after block 17 for the fitted plane. The top-2 PCA-plane control produces substantially weaker steering. **d)** Displacement-match rates across rotation angles, showing three plateaus at -1, 0 and $+1$ with transitions near $\pm 70°$ rather than a gradual effect. **e)** Rotating one internal coding exon produces localized steering. Intronic and displacement-matched random-plane interventions remain near the plotted baseline.

rotations substantially change the decoded amino acid sequence despite a one-nucleotide change in reading frame.

Figure B.1 reports displacement matches (the fraction of translated amino acids that align with the expected reading frame, see A.4) for single rotations, compositions, insertions, and controls. Single rotations of $\pm 120°$ yield approximately 90% floor-corrected match rates at the expected displacement. Scaling and $5'$ UTR rotation preserve the original frame, while the random-plane control degrades the translated protein with an original-displacement match rate of approximately one half. Figure 4e additionally shows that rotating one coding exon localizes steering to that exon.

After a one-nucleotide insertion, the original-displacement match rate is approximately 66%. A three-nucleotide insertion preserves frame modulo three, and subsequent rotations have a similar effect to the no-insertion case. Three successive $120°$ rotations applied across encoder depths also largely restore the original decoded reading frame (Fig. B.1). This behavior is consistent with an action of the cyclic group of order three, C_3, in which rotations cycle the decoder through the three reading frames. The observed recovery provides behavioral evidence that transformations in the learned codon phase plane induce the expected cyclic composition of reading frame shifts.

6 Tracing the Translation Mechanism

We next trace how MIMIC converts codon-level RNA representations into amino acid predictions. We use a codon-substitution assay, where at one interior position of each transcript, we replace the codon with each of the 61 sense (non-stop) codons and ask the model to decode the corresponding residue, either with only that residue hidden or with the entire protein hidden. Four transcripts were used for discovery, and twelve separately reserved transcripts were evaluated for confirmation

under a fixed protocol. The native-readout and decoder-intervention results use the twelve confirmation transcripts (732 codon variants). Linear probes use a separate twelve-transcript cohort (Appendix A.5). Details are provided in Appendix A.5 and A.6.

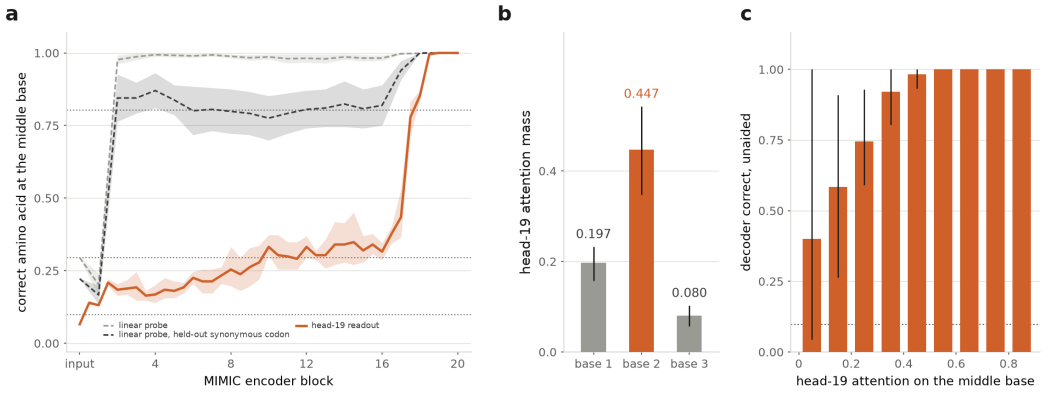

Figure 5: **Native decoder readout and codon addressing in MIMIC. a)** Frozen native readout on twelve confirmation transcripts (61 codons each), compared with linear probes on a separate twelve-transcript cohort and partial-codon lookup baselines. The synonymous-codon probe band spans ridge and LDA estimates. **b)** Mean attention from decoder block 12, head 19 (one-based) to each codon base, with one residue hidden. **c)** Observational association between native accuracy and middle-base attention.

To measure what MIMIC's decoder can read from each encoder position, we compose the weights of a single decoder cross-attention head into a fixed linear readout. For head h with output projection $W_{O,h}$, the amino acid scores obtained if the head attended entirely to encoder position p are

$$s_p = U\,\Gamma\,P\,W_{O,h}\,v_p, \tag{1}$$

where v_p is the value the head retrieves from position p (a fixed affine function of the normalized encoder output there), Γ and P are the gain and centering of the decoder's final layer normalization, and U contains the 20 amino acid rows of the protein output embedding. This readout uses only MIMIC's weights and fits no parameters to amino acid labels. It therefore asks whether the amino acid is present in the channel the decoder reads, rather than whether any classifier could recover it. We apply it to head 19 of decoder block 12, the final decoder block, which discovery experiments identified as the main cross-attention channel carrying amino acid identity (see Fig. A.3).

Fig. 5a shows a sharp climb in decoding accuracy, climbing from 51.0% after block 17 to 86.7% after block 18, 99.7% after block 19's attention update, and 100.0% after its feed-forward update. This late onset is not simply a limit on what is decodable. A linear probe identifies the codon at the middle nucleotide with 96–100% accuracy from encoder block 2 onward, so a trained classifier can learn the genetic-code table itself from early layers. When a ridge probe must generalize to a held-out synonymous codon (one encoding the same amino acid), accuracy remains at 70–81% through block 16 and reaches 100% at block 18. The LDA probe has a higher early plateau (85–93%) but a similar late rise. The rise after block 18 coincides with the decline in steering efficacy after that block (Section 5). Phase remains decodable at later depths, and the intervention identifies a change in its downstream influence rather than disappearance of phase information.

With one residue hidden, head 19 concentrates its attention on the target codon, and within the codon on its middle nucleotide (Fig. 5b). The middle base receives 0.45 of the head's attention, compared with 0.20 and 0.08 for the first and third bases, and is the most-attended base of the codon in 91% of cases. While this may be due in part to MIMIC's bidirectional attention centering on the middle base, there are biological priors that might bias the second codon to be most informative as well. The second codon position is especially informative for amino acid identity and physicochemical properties (Haig & Hurst, 1991). Substitutions at the first and particularly third positions can be synonymous, while changes at the second position always alter the encoded amino acid (Crick et al., 1961). Errors additionally track where head 19 attends. Native accuracy increases monotonically

with attention on the middle nucleotide (Fig. 5c). With one residue hidden, it rises from 40% when this attention is below 0.1 to 99.5% when it exceeds 0.4, and wrong predictions place a mean of 0.18 attention on the middle base against 0.48 for correct ones. Correcting the address removes every wrong-residue error in these assays. Together, these results identify a pathway in MIMIC. Amino acid identity becomes accessible to the decoder's native readout late in the encoder, and a decoder cross-attention head retrieves it preferentially from the codon's middle nucleotide.

7 Composing latent rotations and multimodal conditioning in POLG/POLGARF

The human *POLG* gene encodes a DNA polymerase enzyme that is essential for mitochondrial DNA replication (Rahman & Copeland, 2019). *POLGARF* is a distinct protein translated from an alternative reading frame that overlaps *POLG*, with a CUG start codon 55 nucleotides upstream of the *POLG* coding sequence (Loughran et al. (2020), Fig. 6a). We use this spliced transcript as a test case to investigate whether latent rotation and multimodal conditioning separately control frame-specific content and protein–RNA alignment. Full experimental details can be found in Appendix A.7.

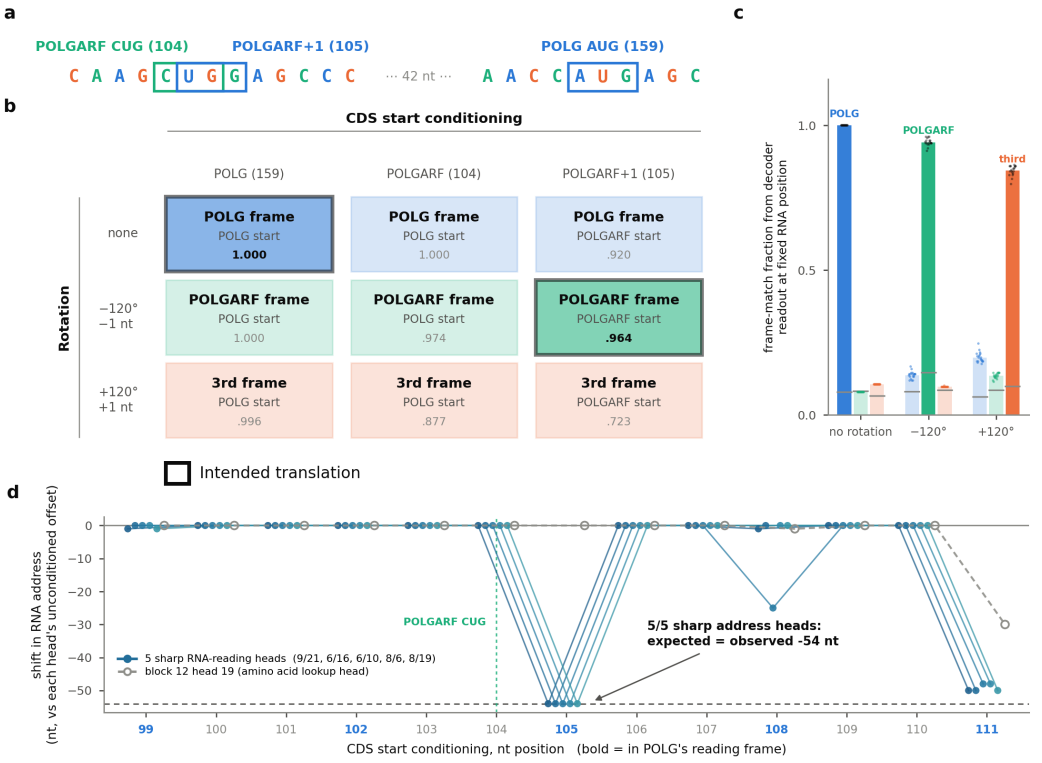

Figure 6: **Frame rotation and start conditioning in POLG/POLGARF.** One transcript, with 15 protein residues supplied and 227 positions scored. **a)** Schematic view of the POLG and POLGARF start codons. **b)** Translated frame and start location for composed rotations and CDS conditioning. **c)** Reading frame readout in MIMIC's decoder at a fixed location. **d)** Median RNA-address shifts relative to each head's unconditioned baseline; dots represent head–condition pairs, with horizontal jitter for visibility. Blue labels mark the original *POLG* codon phase.

First, we attempt to induce a POLG to POLGARF translation change using the same rotation intervention described in Section 5. We find that while the rotations succeed in changing the local reading frame of the decoded protein (see Fig. 6b, first column), since POLGARF begins 55 nucleotides before the canonical strong POLG start codon, the model does not change the address it reads at. We additionally condition MIMIC on the CDS start of POLGARF, but this has a similar

output, only degrading the local frame rotation output (see Fig. 6b, middle column). Remarkably, we find that conditioning the model on a CDS start *one nucleotide after* POLGARF biases the model to read starting at that position (Fig. 6b, right column). A subsequent rotation gives the POLGARF protein output. Results for no CDS conditioning and random rotations, as well as partial decoded amino acid sequences can be found in Fig. B.2.

In the same method as in Section 6 above, we read the amino acid at a fixed RNA position using a readout built only from MIMIC's own weights. Across all 60 combinations tested (three rotations crossed with 20 start codon conditions) the frame we recover is always the one the rotation selects (Fig. 6c). How many residues match varies between annotations, but the decoded reading frame is always faithful to the rotation operation, and is not biased by any CDS conditioning. The annotation can therefore move where the decoder reads without changing the frame the rotation selects.

Finally, we investigate how the CDS conditioning offset impacts the decoder addressing. Fig. 6d shows that register translation is only possible in the native reading frame (a multiple of 3 nucleotide offset from the original CDS start). Notably, the operation is not valid at every multiple of 3. We hypothesize that the model has learned to accept conditioning moving towards only feasible alternative start sites, but further experiments are required to explore this possibility.

While a full investigation of the mechanism of reading frame displacement from multimodal conditioning is left to a future study, we show here how it can compose with our well-described latent codon phase translations to select an alternate reading frame and resulting translated protein. We validate these causal interventions on a real human gene and show that they can recover a biologically occurring overlapping reading frame. Together, these results suggest that the learned codon phase representation is not merely descriptive, but can be manipulated compositionally to control biologically meaningful translation outcomes.

8 Discussion

Our results trace a sequence of representations and computations underlying RNA-to-protein translation in MIMIC. Across genomic foundation models, coding sequence develops strong period-3 structure, and in MIMIC, Evo 2 and NTv3 the dominant representation reorganizes around codon phase. Rotations within this learned phase plane causally induce signed changes in the reading frame used for protein decoding in MIMIC. Late in the encoder, this steering effect diminishes as amino acid identity becomes accessible through the decoder's cross-attention channel, which preferentially retrieves information from the middle nucleotide of each codon. Together, these results suggest a progression from a frame-dependent coordinate system for grouping nucleotides into codons to an amino acid representation accessible through the native decoder pathway.

This progression highlights that information may be linearly recoverable well before it is organized in a form used by downstream computation. The presence of reading frame information across several genomic foundation models therefore does not establish that those models use it similarly. Combining representational analysis with targeted intervention and tracing through the model's own downstream weights provides stronger evidence for mechanism than decodability alone.

Several aspects of this mechanism remain model- and context-dependent. MIMIC is uniquely suited to these causal experiments because it was jointly trained on RNA and protein and provides a native RNA-to-protein decoder. Although one final-layer cross-attention head provides a particularly direct route from the encoder's middle-nucleotide representation to protein output, alternative routes can compensate in less constrained decoding settings. We therefore interpret this pathway as a prominent MIMIC mechanism rather than a unique circuit required for translation.

Biological examples can provide natural tests of the representations found here. Programmed ribosomal frameshifts change the active translation frame at a defined sequence position without a corresponding nucleotide insertion or deletion (Hill & Brierley, 2023). This setting offers an opportunity to test how closely the learned phase representation tracks biologically active reading frames. Another open question is the origin of weak period-3 structure in non-coding regions, which may reflect biological sequence structure, proximity to coding regions, or periodicity inherited from the training distribution. More broadly, these results suggest that biologically meaningful symmetries can emerge as causally active internal structure in genomic foundation models, providing a route from representation analysis toward mechanistic understanding of learned biological computation.

AI USE STATEMENT

In this work, we used generative AI tools for coding, brainstorming, literature search and drafting portions of text. We have reviewed all AI-assisted work. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

Experimental methods and statistical conventions are provided in Appendix A; supplementary analyses are in Appendix B. Anonymized repo at `https://anonymous.4open.science/r/codon-rep-145D/`

REFERENCES

Etowah Adams, Liam Bai, Minji Lee, Yiyang Yu, and Mohammed Alquraishi. From mechanistic interpretability to mechanistic biology: Training, evaluating, and interpreting sparse autoencoders on protein language models. In Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu (eds.), *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 460–476. PMLR, 13–19 Jul 2025. URL https://proceedings.mlr.press/v267/adams25a.html.

Loubna Ben Allal, Qiuyi Li, Maurizio Fiusco, Lewis Tunstall, Kashif Rasul, Ed Beeching, Dana Aubakirova, Carlos Patiño, Thibaud Frere, Anton Lozhkov, Georgia Channing, Thomas Wolf, Diego di Bernardo, and Leandro von Werra. Carbon: Decoding the language of life. *bioRxiv*, 2026. doi: 10.64898/2026.05.22.727119. URL https://doi.org/10.64898/2026.05.22.727119. Preprint.

Sam Boshar, Benjamin Evans, Ziqi Tang, Armand Picard, Yanis Adel, Franziska K. Lorbeer, Chandana Rajesh, Tristan Karch, Shawn Sidbon, David Emms, Javier Mendoza-Revilla, Fatimah Al-Ani, Evan Seitz, Yair Schiff, Yohan Bornachot, Ariana Hernandez, Marie Lopez, Alexandre Laterre, Karim Beguir, Peter Koo, Volodymyr Kuleshov, Alexander Stark, Bernardo P. de Almeida, and Thomas Pierrot. A foundational model for joint sequence-function multi-species modeling at scale for long-range genomic prediction. *bioRxiv*, 2025. doi: 10.64898/2025.12.22.695963. URL https://doi.org/10.64898/2025.12.22.695963. Preprint.

G. Brixi, M. G. Durrant, J. Ku, 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. URL https://doi.org/10.1038/s41586-026-10176-5.

Salvatore Candido, Thomas Hayes, Alexander Derry, Roshan Rao, Zeming Lin, Robert Verkuil, Bryan Z. Wu, Jin Sub Lee, Elise S. Bruguera, Jehan A. Keval, Mykhailo Kopylov, John E. Pak, Wesley Wu, Neil Thomas, Samson Mataraso, Alvin Hsu, Ashton C. Trotman-Grant, Kilian Fatras, Allan dos Santos Costa, Rohil Badkundri, Halil Akın, Deniz Oktay, Jonathan Deaton, Elizabeth Montabana, Hrishita Sitwala, Yue Yu, Marius Wiggert, Dylan Alexander Carlin, Anthony W. Goering, Tomasz Blazejewski, McCullen Sandora, Michael Hla, Tina Z. Jia, Leon H. Kloker, Nicholas J. Sofroniew, Masatoshi Uehara, Jassi Pannu, Sharrol Bachas, Daniel S. Liu, Tom Sercu, and Alexander Rives. Language modeling materializes a world model of protein biology. *bioRxiv*, 2026. doi: 10.64898/2026.06.03.729735. URL https://www.biorxiv.org/content/early/2026/06/04/2026.06.03.729735.

Bilal Chughtai, Lawrence Chan, and Neel Nanda. A toy model of universality: Reverse engineering how networks learn group operations. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (eds.), *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 6243–6267. PMLR, 23–29 Jul 2023. URL https://proceedings.mlr.press/v202/chughtai23a.html.

Francis H. C. Crick, Leslie Barnett, Sydney Brenner, and R. J. Watts-Tobin. General nature of the genetic code for proteins. *Nature*, 192:1227–1232, 1961. doi: 10.1038/1921227a0. URL https://doi.org/10.1038/1921227a0.

Joshua Engels, Eric J. Michaud, Isaac Liao, Wes Gurnee, and Max Tegmark. Not all language model features are one-dimensionally linear, 2025. URL https://arxiv.org/abs/2405.14860.

James W. Fickett. Recognition of protein coding regions in DNA sequences. *Nucleic Acids Research*, 10(17):5303–5318, 1982. doi: 10.1093/nar/10.17.5303.

Siavash Golkar, Alberto Bietti, Mariel Pettee, Michael Eickenberg, Miles Cranmer, Keiya Hirashima, Geraud Krawezik, Nicholas Lourie, Michael McCabe, Rudy Morel, Ruben Ohana, Liam Holden Parker, Bruno Régaldo-Saint Blancard, Kyunghyun Cho, and Shirley Ho. Contextual counting: A mechanistic study of transformers on a quantitative task, 2024. URL https://arxiv.org/abs/2406.02585.

Siavash Golkar, Jake Kovalic, Irina Espejo Morales, Samuel Sledzieski, Minhuan Li, Ksenia Sokolova, Geraud Krawezik, Alberto Bietti, Claudia Skok Gibbs, Roman Klypa, Shengwei Xiong, Francois Lanusse, Liam Parker, Kyunghyun Cho, Miles Cranmer, Tom Hehir, Michael McCabe, Lucas Meyer, Rudy Morel, Payel Mukhopadhyay, Mariel Pettee, Helen Qu, Jeff Shen, David Fouhey, Hadi Sotoudeh, Vikram Mulligan, Pilar Cossio, Sonya M. Hanson, Alisha N. Jones, Olga G. Troyanskaya, and Shirley Ho. MIMIC: A generative multimodal foundation model for biomolecules, 2026. URL https://arxiv.org/abs/2604.24506.

Wes Gurnee, Emmanuel Ameisen, Isaac Kauvar, Julius Tarng, Adam Pearce, Chris Olah, and Joshua Batson. When models manipulate manifolds: The geometry of a counting task. *Transformer Circuits Thread*, 2025. URL https://transformer-circuits.pub/2025/linebreaks/index.html.

David Haig and Laurence D. Hurst. A quantitative measure of error minimization in the genetic code. *Journal of Molecular Evolution*, 33(5):412–417, 1991. doi: 10.1007/BF02103132.

Chris H. Hill and Ian Brierley. Structural and functional insights into viral programmed ribosomal frameshifting. *Annual Review of Virology*, 10(1):217–242, 2023. doi: 10.1146/annurev-virology-111821-120646.

Gary Loughran, Alexander V. Zhdanov, Maria S. Mikhaylova, Fedor N. Rozov, Petr N. Datskevich, Sergey I. Kovalchuk, Marina V. Serebryakova, Stephen J. Kiniry, Audrey M. Michel, Patrick B. F. O'Connor, Dmitri B. Papkovsky, John F. Atkins, Pavel V. Baranov, Ivan N. Shatsky, and Dmitry E. Andreev. Unusually efficient CUG initiation of an overlapping reading frame in POLG mRNA yields novel protein POLGARF. *Proceedings of the National Academy of Sciences*, 117 (40):24936–24946, 2020. doi: 10.1073/pnas.2001433117.

Kevin Lu, Jannik Brinkmann, Stefan Huber, Aaron Mueller, Yonatan Belinkov, David Bau, and Chris Wendler. Two stages of folding: Convergent mechanisms in AI protein folding trunks, 2026. URL https://arxiv.org/abs/2602.06020.

Neel Nanda, Lawrence Chan, Tom Lieberum, Jess Smith, and Jacob Steinhardt. Progress measures for grokking via mechanistic interpretability, 2023. URL https://arxiv.org/abs/2301.05217.

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 Chris Ré. HyenaDNA: Long-range genomic sequence modeling at single nucleotide resolution, 2023. URL https://arxiv.org/abs/2306.15794.

Yuan-Fei Pan, Yong He, Yu-Qi Liu, Yong-Tao Shan, Shu-Ning Liu, Xue Liu, Xiaoyun Pan, Yinqi Bai, Zan Xu, Zheng Wang, Jieping Ye, Edward C. Holmes, Bo Li, Yao-Qing Chen, Zhao-Rong Li, and Mang Shi. Predicting the evolutionary and functional landscapes of viruses with a unified nucleotide-protein language model: LucaVirus. *bioRxiv*, 2025. doi: 10.1101/2025.06.14.659722. URL https://www.biorxiv.org/content/early/2025/06/20/2025.06.14.659722.

Shamima Rahman and William C. Copeland. POLG-related disorders and their neurological manifestations. *Nature Reviews Neurology*, 15(1):40–52, 2019. doi: 10.1038/s41582-018-0101-0.

Arnav Shah, Junzhe Li, Parsa Idehpour, Adibvafa Fallahpour, Brandon Wang, Sukjun Hwang, Bo Wang, Patrick D. Hsu, Hani Goodarzi, and Albert Gu. dnaHNet: A scalable and hierarchical foundation model for genomic sequence learning, 2026. URL https://arxiv.org/abs/2602.10603.

Shrish Tiwari, S. Ramachandran, Alok Bhattacharya, Sudha Bhattacharya, and Ramakrishna Ramaswamy. Prediction of probable genes by Fourier analysis of genomic sequences. *Bioinformatics*, 13(3):263–270, 1997. doi: 10.1093/bioinformatics/13.3.263. URL https://doi.org/10.1093/bioinformatics/13.3.263.

A METHODS

A.1 DATASET CONSTRUCTION

Our cross-model analyses use two complementary transcript cohorts. The microbial cohort contains 297 translation-verified, intronless transcripts from *E. coli*, *S. cerevisiae*, and *M. tuberculosis*. Coding regions are defined by the annotated CDS (coding sequence) regions and non-coding regions by untranslated regions from the same transcripts. The metazoan cohort contains 270 unspliced pre-mRNAs from *D. melanogaster*, *C. elegans*, macaque, and human. Here, coding regions are drawn from coding exons and non-coding regions from internal introns. We retain sufficiently long protein-coding transcripts and sequence segments to support matched position-wise analyses.

The microbial and metazoan cohorts are treated as distinct substrates throughout the analysis rather than pooling all nucleotide positions into a single dataset. The aggregation procedure used to combine them depends on the analysis and is specified in the corresponding methods subsection. Additional spliced metazoan mRNAs spanning Vertebrata, Arthropoda, and Nematoda are used for MIMIC causal translation experiments.

A.2 PERIODICITY IN REPRESENTATION NORMS

For each model and hidden layer, we take the magnitude of the hidden representation at each token index i,

$$x_i = \|h_i\|_2,$$

producing a one-dimensional signal along the sequence. We remove slowly varying trends by subtracting a centered moving-average baseline, using a window width that is an integer multiple of three so that detrending leaves the $f = 1/3$ component unattenuated. For the cross-model analysis, we use a 51-nt window. We then apply a Hann window to reduce spectral leakage and quantify the component at $f = 1/3$ cycles per nucleotide, corresponding to a three-nucleotide period.

Coding vs. non-coding periodicity across models: For the cross-model comparison in Fig. 2c, we estimate spectral power using the periodogram, computed as the squared magnitude of the discrete Fourier transform of the signal described above. Each segment's spectrum is normalized by its median broadband power ($f > 0.02$), so the resulting statistic measures the prominence of the three-nucleotide component relative to that segment's own spectral background. Figures 2b–c use the single $f = 1/3$ bin, distinct from a band-to-flanking-band ratio that integrates frequencies within ± 0.03 of $1/3$.

Within each substrate, we compute a normalized periodogram for every segment and take the pointwise median across segments to obtain a representative spectrum. For the microbial cohort, coding CDS segments are compared with UTRs from the same transcripts. For the metazoan cohort, coding exons are compared with internal introns. The microbial and metazoan estimates are then combined using their geometric mean, giving the two substrates equal weight. For each model, we report the layer with the largest coding-to-non-coding period-3 enrichment.

The reported values in Figure 2c are read at the layer where its coding versus non-coding contrast is largest. Figure A.1 plots the coding to non-coding period-3 ratio at every recorded depth for all four models, with the selected layer for Figure 2c marked with a circle.

Periodicity across RNA classes. The RNA-class comparison in Fig. 2d spans substantially different sequence lengths, including long non-coding RNA (lncRNA), miscellaneous RNA (miscRNA), and ribosomal RNA (rRNA) segments. In this regime, periodogram power is strongly length-dependent, so we instead use a length-normalized Fourier amplitude at the period-3 frequency,

$$A_{1/3} = \left| \frac{2}{N} \sum_{t=1}^{N} x_t e^{-2\pi i t/3} \right|.$$

Before computing the amplitude, each profile is z-scored and high-pass filtered by subtracting a centered 9-nt moving average. As above, the moving-average width is chosen to be a multiple of three so that the $f = 1/3$ component is not attenuated. Statistics are first summarized within genes and then aggregated within RNA class.

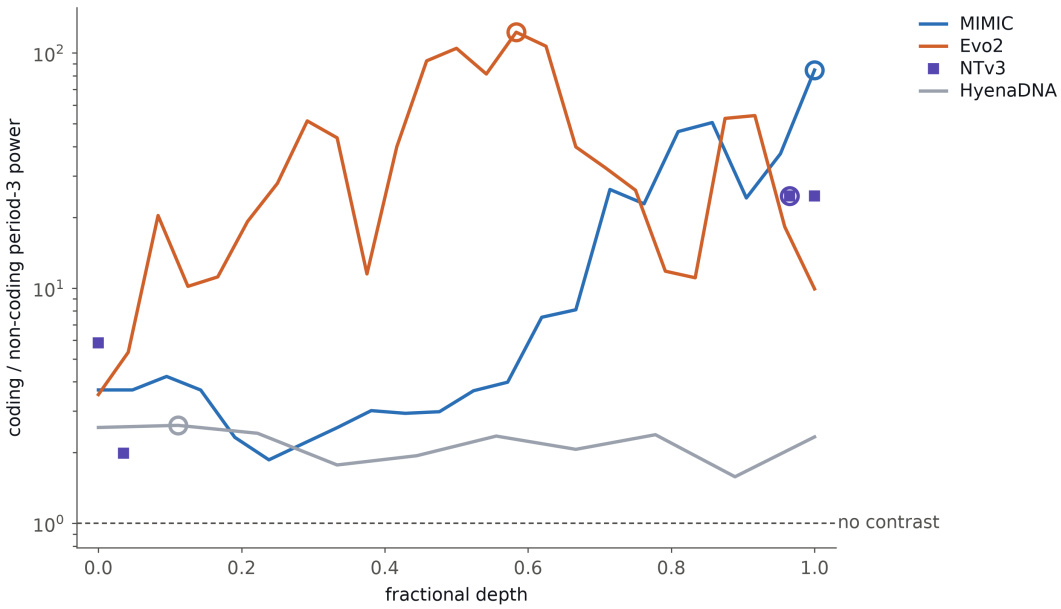

Figure A.1: **Coding / non-coding period-3 power across layers of all models tested.**

Each RNA class is compared against a local non-coding reference determined by its transcript structure. For protein-coding genes, CDS segments are compared with intronic sequence within the CDS span together with available flanking sequence. Multi-exon pseudogenes and lncRNAs are compared with their introns and available flanks. MiscRNA and rRNA are measured over annotated segments, rather than entire padded gene windows. Transfer RNA (tRNA) is excluded from Fig. 2d: only 41 genes contribute short, approximately 69-nt segments, and their flanking reference has stronger period-3 amplitude than the annotated segments, preventing a clean local comparison. UTR exons are excluded from both arms.

Choice of spectral estimator. The power and amplitude estimators have complementary failure modes. Periodogram power provides greater dynamic range for long, strongly periodic coding segments, whereas its coding-to-background ratio varies substantially with segment length. The normalized amplitude is considerably more stable across the heterogeneous segment lengths in the RNA-class analysis, but approaches a ceiling for strongly periodic coding representations and therefore compresses cross-model differences. We therefore use periodogram power for Fig. 2b–c and Fourier amplitude for Fig. 2d.

The alternative estimators preserve the separation of protein-coding genes and pseudogenes from the other displayed RNA classes. Moving-average-width sweeps show that the rRNA estimate changes at the largest widths, where very few rRNA segments remain. Length dependence, estimator comparisons, and retained coverage are shown in Fig. A.2.

A.3 Representational geometry and linear probes

Codon phase: For a nucleotide at position i, we define the phase coordinate

$$\phi_i = (i - i_{\text{CDS start}}) \bmod 3, \tag{2}$$

with positions referenced to the annotated CDS start on spliced input. On unspliced pre-mRNA, coding phase is computed by accumulating CDS-piece lengths in transcript order, including across exon junctions, rather than subtracting raw genomic positions. Each preceding piece contributes its full annotated length even if extraction truncates that piece. Within coding sequence, $\phi_i \in \{0, 1, 2\}$ corresponds to the nucleotide's position within the translated codon. For non-coding regions, the control label is the arithmetic continuation $(\text{pos} - \text{CDS start}) \bmod 3$ in raw input coordinates, as if

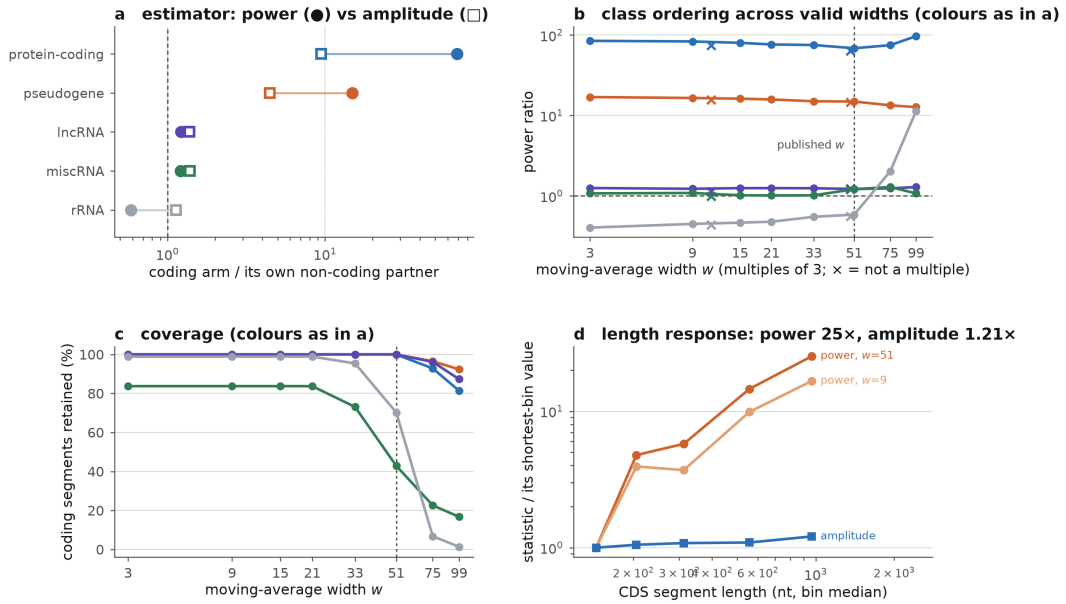

Figure A.2: **Sensitivity of period-3 enrichment to estimator, smoothing width, and segment length. a)** Power and amplitude ratios for the displayed RNA classes, each relative to its local non-coding reference. **b)** Power ratios across moving-average widths; crosses indicate widths not divisible by three. **c)** Retained segment coverage across widths. **d)** Length-binned power and amplitude statistics, each normalized to its shortest-bin value. Large widths substantially reduce coverage for short RNA classes.

the coding register extended through the surrounding sequence. Each sampled transcript is checked to contain more than one control phase. This extrapolated label is used only to test whether the representation organizes around the same periodic coordinate outside coding regions.

Principal-component geometry: To characterize the dominant geometry of the hidden representations, we analyze coding and non-coding nucleotide tokens separately at every recorded layer. Within each token pool, we center the hidden states and fit a two-component PCA. The projection is fully unsupervised and neither nucleotide identity nor codon phase labels are used to fit the principal components.

The non-coding pool in Fig. 3 combines five UTR subsets (microbial 5′, human 5′ and 3′, and metazoan 5′ and 3′) with an organism-stratified pre-mRNA intron subset, balanced by record count across subsets. The intron subset is metazoan-only and cannot be balanced against the intronless microbial substrate.

After fitting the projection, we quantify how strongly nucleotide identity and codon phase organize the resulting two-dimensional geometry using

$$\eta^2 = \frac{\sum_k n_k \|\bar{z}_k - \bar{z}\|^2}{\sum_i \|z_i - \bar{z}\|^2}, \tag{3}$$

where n_k is the number of tokens in class k, z_i is the two-dimensional PCA representation of token i, $\bar{z}_k$ is the centroid of class k, and $\bar{z}$ is the global mean. η^2 is therefore the fraction of variance in the leading two-dimensional representation associated with the class labels.

Linear probes: We separately measure whether nucleotide identity and codon phase are linearly decodable from higher dimensional hidden representations. At each layer, activations are reduced to 150 principal components and evaluated using five-fold cross-validation grouped by transcript, so that tokens from the same transcript never appear in both training and test folds. The unlabeled PCA basis is fitted once using all folds; test-fold covariates therefore contribute to the projection,

although test labels do not. Features are standardized using the training fold. We fit multinomial logistic regression and a shrinkage linear discriminant classifier as independent linear readouts and verify that the two give consistent conclusions. Chance accuracy is $1/4$ for nucleotide identity and $1/3$ for codon phase.

The PCA and probe analyses answer complementary questions: η^2 measures whether a variable organizes the model's *dominant* low-dimensional geometry, whereas probe accuracy measures whether the same information is linearly recoverable anywhere in the representation.

A.4 Causal Steering

Fitted codon phase plane: At each layer of MIMIC's encoder, we fit a two-dimensional codon phase subspace using a held-out transcript set. We first compute one centroid for each codon phase, μ_0, μ_1, μ_2, by averaging the residual-stream representations of tokens with that phase. We then subtract the mean representation over all fitted tokens from each centroid and define the frame plane as the two-dimensional subspace they span. The resulting orthonormal basis $U \in \mathbb{R}^{d\times 2}$ is oriented so that phases $0 \to 1 \to 2$ proceed counter-clockwise, fixing the sign convention such that a $+120^\circ$ rotation corresponds to a $+1$ phase change.

This intervention plane is distinct from the unsupervised PCA projection used to characterize representational geometry above. PCA identifies directions of maximal variance, whereas the frame plane isolates the two-dimensional subspace associated specifically with codon phase. The PCA plane is evaluated separately as an intervention control to test whether steering depends specifically on the phase-aligned subspace rather than on any dominant low-dimensional direction.

Intervention and readout: For a hidden state h at the intervention layer, we rotate only its component within the fitted frame plane:

$$h' = h + U(R_\theta - I)U^\top(h - c), \tag{4}$$

where c is the transcript-specific mean over the same set of tokens being edited and R_θ is a two-dimensional rotation matrix. Components orthogonal to U are unchanged. Our primary interventions use $\theta = \pm 120^\circ$, corresponding to the two signed changes in codon phase. Unless otherwise stated, the same transformation is applied uniformly to all nucleotide tokens within the CDS at a single encoder layer, and we additionally sweep the intervention across encoder depth.

We evaluate the intervention through MIMIC's native RNA-to-protein decoder. At a fixed protein output position, the decoded residue is compared with the amino acid specified by the original reading frame and by the $+1$ and -1 shifted reading frames. This readout is position-preserving, since the decoder output position is held fixed while the nucleotide triplet associated with that position changes under the intervention. Thus the assay tests whether rotating the latent frame representation produces the corresponding signed change in the model's protein output.

Steering statistic and evaluation cohort: Evaluation is restricted to transcripts whose unmodified top-1 residue accuracy is at least 0.99: 227 of 227 microbial and 162 of 208 metazoan transcripts pass, giving 389 evaluated transcripts. At each candidate nucleotide displacement o, residue matches and scored positions are pooled within each substrate, and the two substrate rates are averaged equally to give $m(o)$. For an expected displacement o_*, the matching floor is the median of $m(o)$ over offsets satisfying $|o - o_*| > 1$ and $(o - o_*) \bmod 3 \neq 0$. The reported statistic is $\max(0, m(o) - \text{floor})$. Exact-offset matching is distinct from frame agreement modulo three; offsets separated by a codon have the same frame. Confidence intervals use 10,000 transcript-bootstrap replicates, stratified by substrate, with seed 0 and paired draws across intervention arms; the floor is re-estimated in every replicate. Intervals cover transcript sampling within these fixed substrates, not model, intervention-seed, or substrate-selection uncertainty.

Insertion operations: To test whether the learned transformation composes with sequence-level changes in reading frame, we additionally insert one, two, or three nucleotides into the RNA sequence before the forward pass. An insertion of k nucleotides shifts the downstream triplet register by $-k$ modulo three. We evaluate sequence insertions alone and in composition with $\pm 120^\circ$ latent rotations, testing whether the effects combine according to the expected modulo-three frame arith-

metic. A one-nucleotide deletion arm exists in the result table but is not plotted in Fig. B.1. We also apply rotations repeatedly to test composition and closure under three successive 120° rotations.

Controls: We use several controls to distinguish frame-specific steering from generic effects of perturbing the residual stream. A displacement-matched random-plane control applies the same rotation in a random two-dimensional subspace orthogonal to the fitted frame plane. An in-plane scaling control changes the magnitude of the frame-plane component by an amount matched to the displacement of a 120° rotation without changing its angular coordinate. We further test rotations of $\theta \neq \pm 120^\circ$, and apply the frame rotation to non-coding 5′UTR tokens as a scope control. Random-plane controls are matched to the number of edited tokens and perturbation magnitude of the corresponding intervention.

A.5 Codon to amino acid computation

Codon-substitution assay. We use sixteen spliced metazoan transcripts with exactly translated annotated coding sequences, protein lengths of 70–220 residues, and RNA lengths of 240–1000 nucleotides. Exact protein duplicates are excluded; homologous proteins are not clustered, and exposure during pretraining is unknown. In each transcript we select one interior target codon that originally encodes arginine, leucine, or serine, and replace it with each of the 61 sense (non-stop) codons, preserving sequence length and reading frame. Each variant is evaluated in two conditions: *one residue hidden*, in which only the target protein residue is masked and all other residues are supplied, and *entire protein hidden*, in which the model receives the RNA and the true protein length but no amino acid identities. Both use MIMIC's native masked-output forward pass in bfloat16, and neither tests free-running generation with inferred start, stop, or length. Four transcripts were used for adaptive discovery, including selection of the decoder head analyzed below. The remaining twelve were reserved, and the confirmation measurements, intervention definitions, random seeds, and analysis were fixed before these transcripts were evaluated. Native-readout and decoder-intervention measurements in Section 6 use the confirmation transcripts; the linear probes use a separate cohort.

Native amino acid readout. MIMIC's decoder attends to a context formed from the encoder output. For encoder position p with final encoder residual x_p, the value that decoder cross-attention head h retrieves is

$$v_p = W_{V,h}\left(W_c\left(\mathrm{LN}_{\mathrm{enc}}(x_p) + e_p\right) + b_c\right), \tag{5}$$

where $\mathrm{LN}_{\mathrm{enc}}$ is the encoder's final normalization, e_p is the modality and position embedding added to the encoder output, (W_c, b_c) is the decoder's context projection, and $W_{V,h} \in \mathbb{R}^{64\times d}$ is the head's value projection. The head's direct contribution to the amino acid logits is obtained by passing its output through its output projection $W_{O,h}$, the centering operator $P = I - \mathbf{1}\mathbf{1}^\top/d$ and gain Γ of the decoder's final layer normalization, and the 20 amino acid rows U of the protein output embedding. We define

$$B^\top = U\,\Gamma\,P\,W_{O,h}, \qquad s_p = B^\top v_p \in \mathbb{R}^{20}. \tag{6}$$

s_p gives the amino acid scores the head would write if it attended entirely to position p, up to division by the final residual standard deviation, which is positive and common to all classes and therefore does not change the predicted amino acid $\arg\max s_p$. All matrices are MIMIC's frozen weights, cast to the native inference precision, and no parameters are fitted to amino acid labels. These scores are direct contributions of one head, not full-model logits: other heads, earlier residual components, and the final decoder feed-forward update also contribute to the output. Throughout, we apply the readout at the middle nucleotide of the target codon, a position fixed before confirmation, and score the prediction against the genetic code across all 61 codons.

Readout across encoder depth. For Fig. 5a, the native-readout measurements use the twelve confirmation transcripts. We truncate the encoder after the feed-forward update of block 17 or 18, or after the attention or feed-forward update of block 19. Every subsequent encoder write is zeroed at all positions, while the native final encoder normalization and full decoder are kept. The readout is applied to the resulting encoder output. The block-20 feed-forward cutoff removes no updates and is the untruncated run on the same cohort. Accuracy is 0.510, 0.867, 0.997, and 1.000 at the four specified late cutoffs, respectively. Thus the cumulative encoder residual already supports near-perfect native-channel readout after block 19's attention update, before its feed-forward update.

This does not establish which earlier component is necessary for that information or when it first becomes recoverable by other readouts. The earlier discovery analysis decomposed cumulative residual writes on four transcripts; that decomposition is not the confirmation readout curve. The two protocols should not be conflated with reading an isolated feed-forward write.

Reference accuracies. The dashed lines in Fig. 5a give the best accuracy achievable over the 61 sense codons by a lookup on partial codon information: predicting the majority amino acid ($6/61 = 9.8\%$), predicting from the middle base alone ($18/61 = 29.5\%$), or predicting from the first two bases ($49/61 = 80.3\%$). All 61 sense codons are weighted equally. For the partial lookups, each value of the partial codon is assigned its most frequent amino acid. These are ceilings for lookups on base identity, not for arbitrary classifiers of the encoder state.

Comparison with linear probes. To distinguish availability in the decoder's channel from linear decodability, we compare the readout with probes fit at the middle nucleotide after every encoder block. The probes use a separate cohort of twelve translation-verified metazoan transcripts, disjoint from the sixteen above, evaluated with the same 61-codon substitution and one residue hidden. At each block, representations are standardized, reduced to at most 64 principal components fit on the training fold, and standardized again. We then fit a ridge classifier ($\alpha = 10$) and, separately, a shrinkage linear discriminant classifier, using three-fold cross-validation grouped by transcript. We use two splits. In the *held-out transcript* split, test transcripts are unseen but every codon appears in training, so a probe can succeed by recognizing the codon and learning the 61-to-20 genetic-code table itself. In the stricter *held-out synonymous codon* split, for each of the 18 amino acids with more than one codon, one codon is removed from training and used only at test time. The probe must then assign the correct amino acid to a codon it has never seen, which requires the representation of that codon to share amino acid-specific structure with its synonyms. A codon-identity probe reaches 96–100% from block 2 onward under both classifiers. On the stricter split, the ridge probe stays between 70% and 81% through block 16, then reaches 91% after block 17 and 100% after block 18. The shrinkage LDA probe gives a higher plateau (85–93%) but the same late rise ($89\% \to 97\% \to 100\%$ over blocks 16–18). The height of the plateau therefore depends on the classifier, while the rise in blocks 17–18 does not. As with the readout, these probes are evaluated in the 61-codon panel and do not characterize the representation of natural sequence in general.

A.6 Decoder readout and codon addressing

Head selection. All analyses target the final cross-attention update of the decoder (block 12 of 12), at the output position of the target residue. On the discovery transcripts, we decompose the decoder's amino acid logits into direct contributions from each residual write using the same readout construction as above, and find that the largest contribution to the correct amino acid score comes from cross-attention head 19 of block 12, with head 21 as a secondary contributor. Both heads were selected on the discovery transcripts, and no head, subspace, or threshold was tuned on the confirmation transcripts. We number blocks and heads from one.

Attention measurements. We recompute head 19's attention weights at the target output position from its captured queries and keys, using the native attention mask and bias; the attention-weighted sum of values reconstructs the native head output to within 1.5% relative error. We group the attention mass into the first, middle, and third nucleotide of the target codon, all other RNA positions, register tokens, and visible protein positions. Fig. 5b reports the mean mass on each nucleotide of the target codon across the 61 codons of each confirmation transcript. We also report the fraction of variants in which the middle nucleotide receives the most attention among the three bases of the codon. With one residue hidden, the three bases receive 0.20, 0.45 and 0.08 of the attention, the codon receives 0.73 in total, and the middle nucleotide is the most attended base in 90.7% of variants. With the entire protein hidden, the corresponding masses are 0.06, 0.18 and 0.14, the codon receives 0.38, and the middle nucleotide is the most attended base in 49.3% of variants. Attention weights describe where the head reads from; they do not by themselves establish that the attended information is used, which the interventions below test.

Head ablation and retention. To test whether head 19 is needed, we set its output to zero at the target output position in the final cross-attention update, zero both heads 19 and 21, or zero every

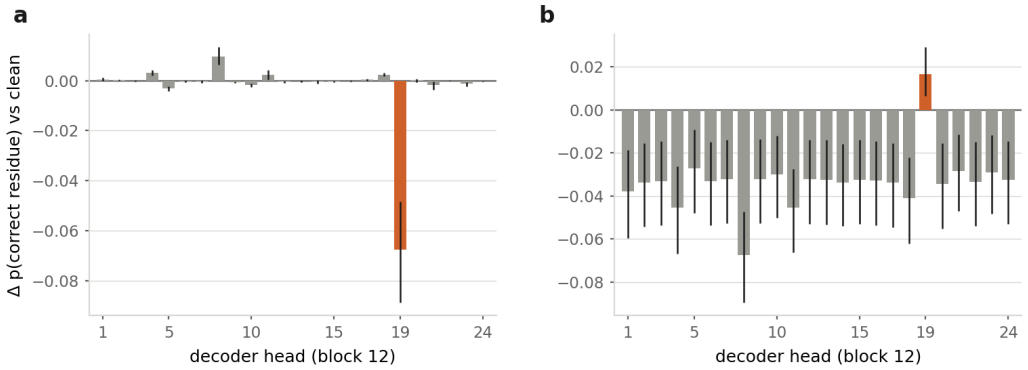

Figure A.3: **Head 19 of the final decoder layer has the greatest impact in amino acid selection.** **a**) Leave-one-out ablation. **b**) Leave-all-others-out ablation.

head except head 19. Earlier decoder computation and all other residual paths are unchanged. With one residue hidden, zeroing head 19 lowers accuracy from 88.5% (95% CI 75.3–99.2%) to 53.4% (35.2–72.4%), zeroing heads 19 and 21 gives 51.4%, and keeping only head 19 gives 87.4%. With the entire protein hidden, none of these interventions changes accuracy appreciably (85.5%, 83.9% and 85.7%, against 85.0% native), indicating redundant routes in that setting. A small ablation effect does not imply that the head is unused.

Score-channel transfer. To test whether head 19's output carries amino acid identity in the directions scored by B, we run a donor variant (target codon edited) and an unedited recipient, and insert the change in head 19's output, $\delta = a_{\text{donor}} - a_{\text{recipient}}$, into the recipient at the target position. We insert either the full change δ, its projection $QQ^\top\delta$ onto the score-contrast subspace, its complement $\delta - QQ^\top\delta$, or a random direction of norm $\|QQ^\top\delta\|$, averaged over three fixed seeds. Here Q is an orthonormal basis for the rank-19 span of the amino acid-centered columns of B, a rank that follows from the 20 output classes rather than being a discovered structure. On nonsynonymous edits with one residue hidden, the full change makes the recipient predict the donor amino acid in 58.0% of cases and the score component alone in 52.1%, while the complement and random controls never do (0.0%). The mean increase in donor-amino acid probability is $+54.1$ and $+48.8$ percentage points for the full change and score component, against $+0.1$ for both controls. Synonymous edits are excluded from this denominator, since preserving the original amino acid would count as success, and native donor errors are retained.

Address interventions. To separate the content of the representation from where the head reads it, we replace head 19's attention-weighted output at the target position with the value v_p at a specified RNA position, leaving the rest of the forward pass native. In the correct-address condition, p is the middle nucleotide of the target codon. In the wrong-address control, p is the middle nucleotide of the following codon, and predictions are scored against both the target and the next codon's amino acid. Addresses are computed from the annotated coding-sequence coordinates and never from the amino acid answer. For whole-protein decoding, we supply the correct middle-nucleotide address at every output position of each unedited confirmation transcript with the entire protein hidden, and compare sequence accuracy with native decoding, counting each transcript once. Because the coordinates are supplied externally, these interventions show that the representation at the correct position is sufficient; they do not show that MIMIC locates the coding sequence unaided.

Attention and errors. For Fig. 5c, we bin confirmation variants by the attention head 19 places on the middle nucleotide, using ten equally spaced bins with edges $0, 0.1, \ldots, 1.0$, and report native accuracy within each bin. Empty bins are omitted; the top bin, $[0.9, 1.0]$, is empty, leaving nine plotted bars. The thinnest populated bin contains 24 variants from three transcripts, and every other populated bin contains at least 30 variants. Intervals are 95% bootstrap intervals resampling the twelve transcripts, never the 732 variants independently. A zero-width interval indicates no between-transcript variation in the resampled statistic, not high precision in an independent-sample

experiment. Both the attention and the prediction are from the unmodified model, so this panel describes an observational association rather than an attention intervention. Native errors are concentrated in a small number of transcripts: three transcripts account for 82 of the 84 errors with one residue hidden, and two account for 103 of the 110 errors with the entire protein hidden. Codon variants within a transcript are therefore not independent, and these bins describe an association between attention and accuracy rather than a causal effect of attention alone; the address interventions above provide the corresponding causal test. With the entire protein hidden, many variants are decoded correctly despite low attention on the middle nucleotide, consistent with the redundant routes indicated by the ablations.

A.7 CDS START CONDITIONING AND POLG/POLGARF INVESTIGATION

Sequence and coordinate conventions. We use a 1014-nt spliced RNA window from human transcript `ENST00000706918.1` (minus-strand locus chr15:89,329,908–89,334,809). Nucleotide offsets are zero based within this window and intervals are half open. The *POLG* AUG occupies $[159, 162)$, the *POLGARF* CUG occupies $[104, 107)$, and an upstream open reading frame starts with AUG at $[21, 24)$. Translation in the *POLG* frame matches the first 285 residues of UniProt P54098 at all 285 positions; the alternative-frame reference contains 242 *POLGARF* residues. A declared start at offset p is represented by a CDS-junction mark at $p - 1$: the natural *POLG* and *POLGARF* marks are 158 and 103, respectively. Figure 6 labels declared-start offsets, not mark offsets.

Conditioning and protein input. The `cds_junctions` modality is a per-position track with classes `0`, `1`, and `2`. Here it is supplied as a visible encoder input, not requested as an output target. Every arm but one exposes exactly *two* visible track positions, with all others masked; the visible count is therefore constant across the set, which is what makes the arms comparable without a per-arm baseline. Each of the thirteen swept arms places class `1` at the swept mark $m \in \{98, \ldots, 110\}$ and non-start class `0` at the original *POLG* mark 158. Four position-matched null arms use marks 101, 103, 104, and 107 with the same visible positions but non-start classes at both. Two whole-locus controls complete the set: an own-start arm asserting class `1` at the *POLG* mark 158 with class `0` at the *POLGARF* mark 103, and a denial arm placing class `0` at both real marks. The no-conditioning arm is the single exception, supplying no visible junction-track positions at all. The own-start arm matches the no-conditioning addresses in all six analyzed heads.

The first $K = 15$ protein residues are supplied as encoder input and residues 15–284 are requested as targets. Every arm scores the same 227 residue positions, indexed 15–241. The -18-residue register shift is absent at $K = 0$ under the same start conditioning; the dependence on the supplied prefix is unresolved. Consequently these experiments do not test generation without protein-sequence conditioning.

Encoder intervention. We transport an orthonormal 1536×2 plane fitted to 70 transcripts from a 278-record metazoan store, using seed 0 and filters requiring at least 30 nt of $5'$ UTR and at least 300 nt of CDS. The fit uses 8400 tokens, with phase counts 2896, 2829, and 2675. The basis is stored as `U_18`; its subscript denotes snapshot 18, corresponding to the post-feed-forward residual of MIMIC encoder block 17 (zero-based sub-block 33). Only RNA rows in $[159, 159 + 3 \times 285)$ are edited. The center is recomputed on each forward pass as the mean of those edited rows; only the basis is transferred from the fit. UTR rows are not edited. Identity mapping reproduces unmodified logits within 10^{-5}.

The stored basis and displayed signed angles use opposite orientations: implementation maps `rot2(120)` and `rot2(240)` induce -1-nt and $+1$-nt frame changes and are displayed as $-120°$ and $+120°$, respectively. The relocated addresses near offset 105 lie upstream of the edited span, which begins at 159.

Content readout and controls. The fixed-position content readout uses the bare value vector from decoder block 12, one-based head 19, before attention mixing, composed with the head output projection, final layer-normalization gain and centering, and amino acid output weights as in Appendix A.5. No readout parameters are fitted. We compare its amino acid predictions with the three frame references for each of 20 conditioning arms and three rotation conditions. In all 60

combinations the highest-matching frame is the one selected by the rotation. Across conditioning arms, match rates for that frame are 1.0000 without rotation, 0.9119–0.9604 for the -1 frame, and 0.7974–0.8634 for the $+1$ frame. This establishes stability of frame identity, not identical residue-level accuracy across conditioning arms.

Chance is computed separately for each arm as $\sum_a f_{\mathrm{pred}}(a) f_{\mathrm{ref}}(a)$, using the empirical predicted and reference amino acid frequencies. For the selected frame it is 0.0794 without rotation, 0.1469–0.1485 for the -1 frame, and 0.0962–0.1001 for the $+1$ frame. The no-conditioning arm reproduces the *POLG* frame and output register at 1.000, identical to supplying its own start. Displacement-matched random-plane controls yield no coherent frame, with match rates 0.352, 0.295, and 0.357 across the three displayed conditioning columns. These two controls are computed but not drawn in Figure 6.

RNA addresses and head selection. For each output slot, an RNA address is the argmax of a head's cross-attention row restricted to RNA key columns. We retain modal addresses and their fractions, but plot the median address across the 227 scored slots. A plotted shift subtracts the corresponding median in that head's own no-conditioning forward pass. Head selection ranks the scan's address-concentration statistic on a one-in-four subsample of slots, after requiring total RNA attention mass of at least 0.50. The five highest-ranking heads are selected; the amino acid readout head is also included. Their identities, in the figure's mixed convention of one-based decoder block and zero-based head, are $(9, 20)$, $(6, 15)$, $(6, 9)$, $(8, 5)$, and $(8, 18)$, plus $(12, 18)$. The five selected heads each put more than 0.99 of their attention mass on RNA. The subsampled scan statistic is used only for ranking and is distinct from the full-cell modal-address fraction. In the rest of the manuscript, $(12, 18)$ is decoder block 12, one-based head 19.

At mark 104 (declared start 105), all five selected heads shift by exactly -54 nt relative to their own no-conditioning baselines, while the readout head does not shift. The output-register displacement is -18 residues, with adoption rate 0.920. The remaining -1-nt rotation supplies the frame change, giving the total -55-nt separation of the natural starts. By contrast, mark 103 (declared start 104, the natural CUG) shifts none of the six heads. All eight off-register marks give zero shift in every head; compatibility is defined relative to the original *POLG* mark 158 and is unchanged by rotation. Of the five compatible marks, 104 relocates cleanly and 110 relocates inconsistently (selected-head medians -48 or -50 nt; register adoption rate 0.465), while 98, 101, and 107 produce no consistent relocation. The same declared triplet, CCA, occurs at rejected mark 98 and partially accepted mark 110, and acceptance is not monotonic in distance from the original start. The mechanism selecting compatible marks remains unresolved.

All POLG results concern one transcript. The 227 scored positions and the multiple heads or conditioning arms are not independent biological replicates. No resampling, confidence intervals, or within-locus significance tests are used. Dots in the address panel are head–condition medians with horizontal jitter, not replicate estimates.

A.8 Statistical analysis and evaluation

Units of analysis and uncertainty. We use the following statistical conventions across the analyses above. Because multiple nucleotide or amino acid positions are sampled from the same transcript, positions are not treated as independent observations for uncertainty estimation. Unless otherwise stated, confidence intervals are 95% bootstrap intervals obtained by resampling transcripts with replacement. For analyses containing both microbial and metazoan substrates, resampling is stratified by substrate and the same bootstrap draw is used across paired experimental arms. Thus, repeated sampling of a transcript carries all positions associated with that transcript into the bootstrap replicate.

Some reported rates are naturally residue-weighted. In the frame-steering experiments, for example, matches and scored positions are pooled within each substrate before computing a rate, and the microbial and metazoan rates are then macro-averaged with equal weight. Bootstrap resampling nevertheless occurs at the transcript level, preserving the within-transcript dependence among scored positions.

Model selection and cross-validation. Linear probes use transcript-grouped cross-validation, as described in Appendix A.3, so that no transcript contributes tokens to both training and test folds. For the cross-transcript steering evaluation, subspace-fitting and evaluation transcripts are disjoint. For the cross-model period-3 comparison in Fig. 2c, confidence intervals are conditional on the selected layer and do not include uncertainty associated with choosing the layer of maximal coding–non-coding enrichment.

Paired comparisons and controls. Whenever two intervention arms are evaluated on the same transcripts, comparisons are paired at the transcript level. Intervention arms are also compared with the corresponding unmodified forward pass from the same checkpoint and run. Random-plane controls are generated relative to the plane used for that intervention and matched in the number of edited tokens and perturbation magnitude where applicable.

Chance levels and excluded positions. Chance levels are defined separately for each assay. Codon phase classification has chance accuracy $1/3$ and nucleotide-identity classification has chance accuracy $1/4$. For the 61-sense-codon assay used in the translation-mechanism analysis, the appropriate constant-amino acid baseline is $6/61$ rather than $1/20$. When scoring shifted reading frames, positions whose corresponding shifted codon is a stop codon do not specify an amino acid and are excluded from the denominator rather than counted as incorrect.

B SUPPLEMENTAL RESULTS

Figure B.1 supplies the displacement-match profiles discussed in Section 5.

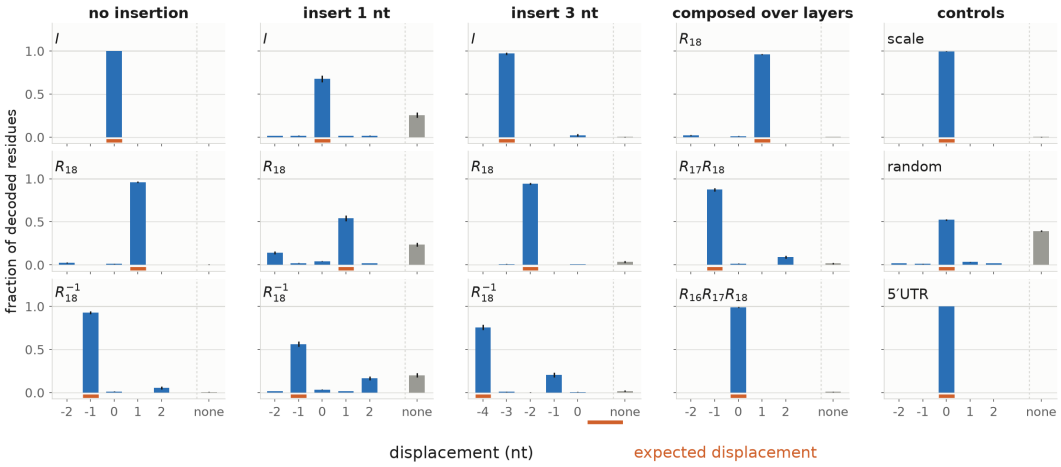

Figure B.1: **Reading frame steering under insertions, compositions, and controls.** Columns compare no insertion, one- and three-nucleotide insertions, successive rotations across MIMIC encoder snapshots, and control interventions. I denotes no rotation; R_ℓ and R_ℓ^{-1} denote opposite 120° rotations at snapshot ℓ, after encoder block $\ell - 1$. Blue bars show floor-corrected displacement-match rates, gray bars show the outcome labeled "none," and orange marks indicate expected displacements. The three-nucleotide insertion preserves frame modulo three while changing absolute sequence coordinates. Rates are macro-averaged across 227 microbial and 162 metazoan transcripts; uncertainty is a 95% stratified transcript bootstrap (10,000 replicates). Intervention and scoring methods are in Appendices A.4 and A.8.



Figure B.2: Decoded proteins for all test cases in the POLG experiments. Black boxes mark the default POLG translation and the successful POLGARF steering translation.

B.1 PROBING OTHER MODELS

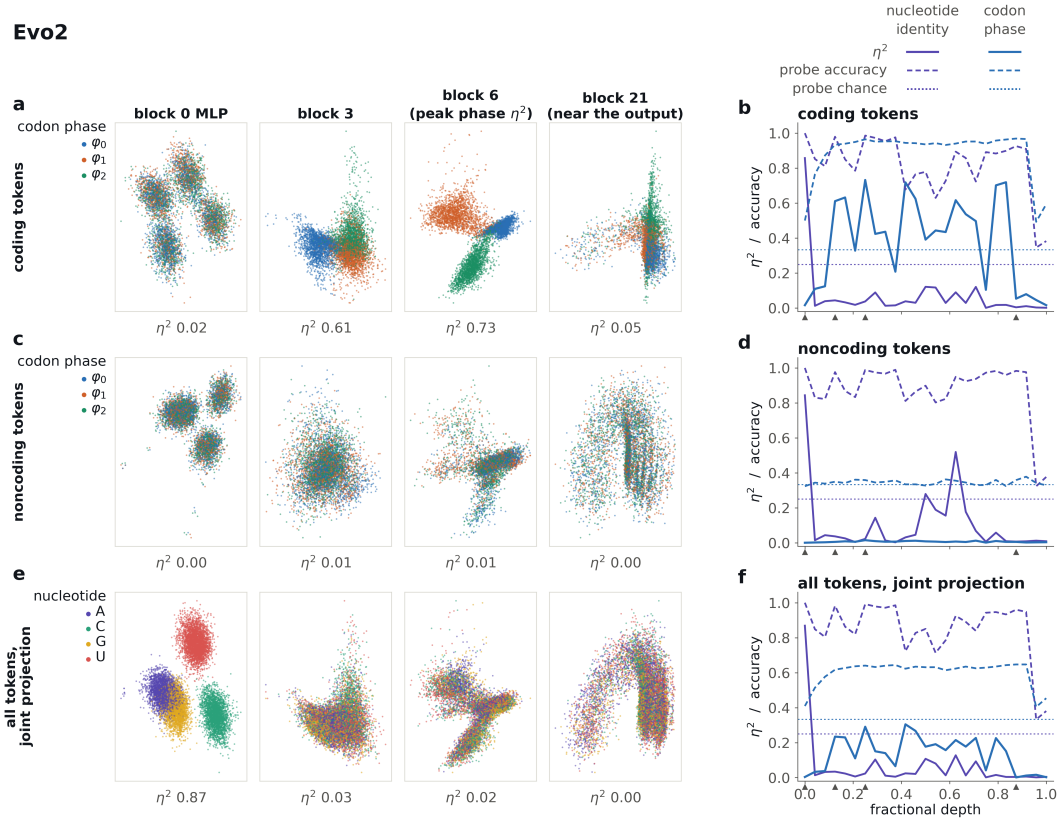

Figure B.3: **Evo 2 builds a strong codon phase plane mid-stack and loses it before its output. a, c, e)** the leading plane at four depths for coding tokens by phase, non-coding tokens by continued phase, and all tokens by nucleotide. **b, d, f)** η^2 and decodability against fractional depth. Chance is $1/3$ for codon phase and $1/4$ for nucleotide.
In Evo 2, codon phase η^2 rises to 0.73 at block 6, with probe accuracy 0.97, while nucleotide η^2 decreases from 0.87 at the first MLP. By block 21, codon phase η^2 decreases to 0.05. Codon phase η^2 remains ≤ 0.02 for non-coding tokens.
Because Evo 2 is autoregressive, early positions have limited left context. We observe activation magnitudes one to two orders of magnitude larger over the first 12 nucleotides, which are therefore excluded from every row of all three appendix figures.

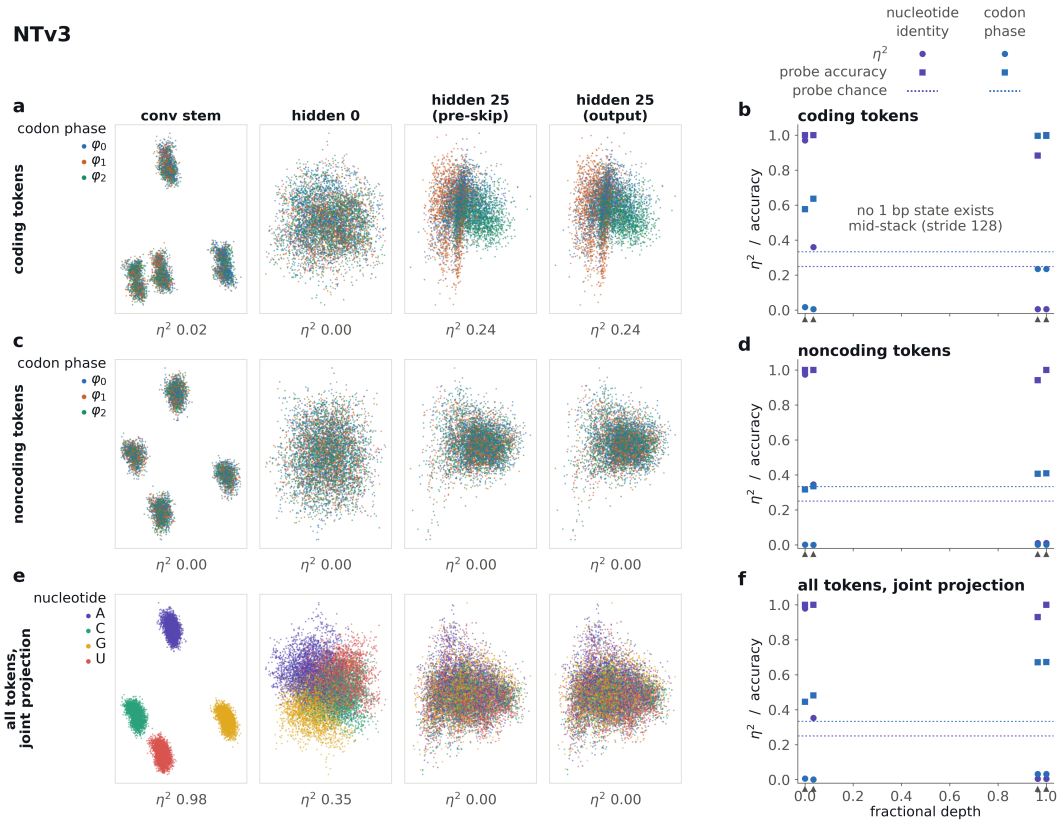

Figure B.4: **Codon phase is detectable at NTv3's single-nucleotide-resolution stem and output.** Layout as in Figure B.3.
NTv3's central architecture is a transformer tower at stride 128 (Boshar et al., 2025), roughly 43 codons per token. A per-nucleotide phase label cannot be evaluated there with the same one-token-per-nucleotide analysis. Only four representations operate at single-nucleotide resolution and all four sit at the two ends of the stack. At the output depths codon phase is decodable at 1.00 on coding tokens and 0.41 on non-coding tokens, the latter close to the $1/3$ chance level. The corresponding codon phase η^2 values in the drawn planes are 0.24 for coding tokens and approximately zero for non-coding tokens.

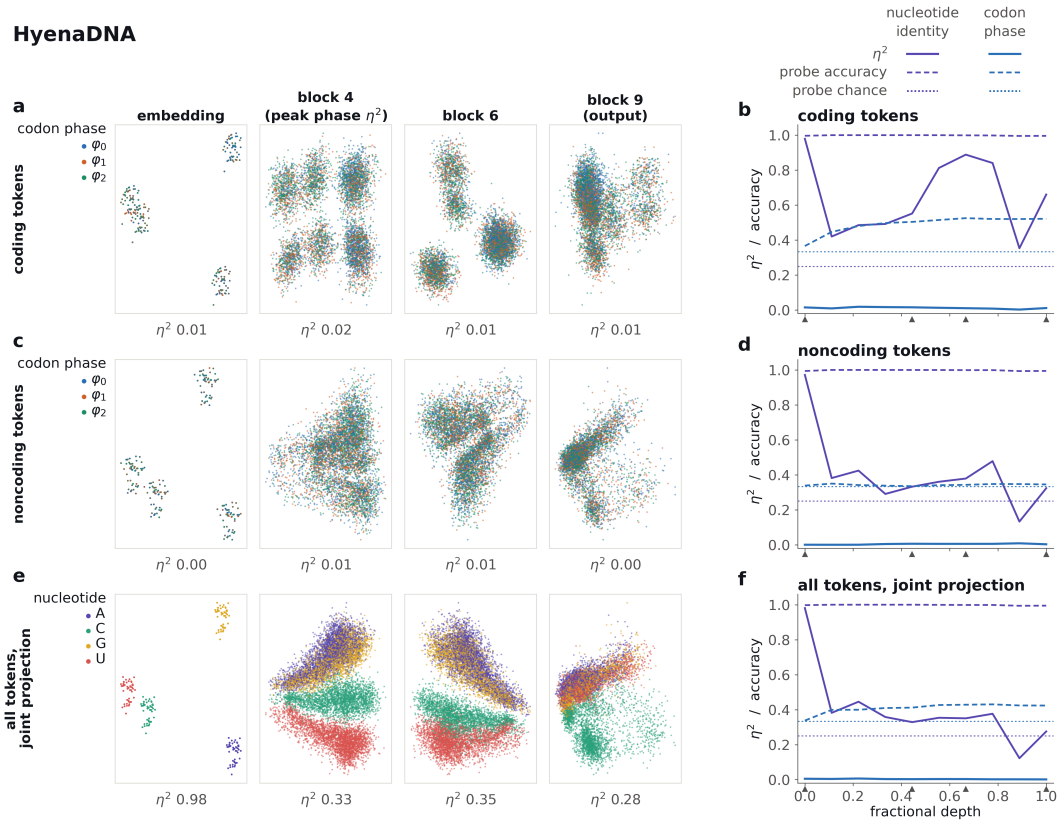

Figure B.5: **HyenaDNA shows weak codon phase organization in its leading representation across depth.** Layout as in Figure B.3.
In HyenaDNA, codon phase η^2 remains at or below 0.02 across coding-token representations, while phase probe accuracy peaks at 0.53. Nucleotide-identity η^2 reaches 0.98 and probe accuracy reaches 1.00. This is consistent with strong preservation of nucleotide identity but comparatively weak codon phase organization, and it matches its 2.6× coding to non-coding period-3 ratio in Figure 2c, the lowest of the four.