Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Findings from Sparse Autoencoders for DNA Sequence Models: Motif Detectors, Reading-Frame Features, and the Scarcity of Regulatory Logic

Olivia Denvis

oliviadenvis@gmail.com

Lampung University https://orcid.org/0009-0002-7443-2669

Research Article

Keywords:

Posted Date: July 22nd, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10434163/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: The authors declare no competing interests.

Findings from Sparse Autoencoders for DNA Sequence Models: Motif Detectors, Reading-Frame Features, and the Scarcity of Regulatory Logic

Olivia Denvis
Independent Researcher
oliviadenvis@gmail.com

Abstract

Sparse autoencoders (SAEs) have become the standard tool for decomposing the internal activations of language models into human-interpretable features, but their behaviour on *DNA sequence models*—transformers and long-convolution models pretrained on genomes—remains largely uncharted. Rather than train an exhaustive suite, we present a deliberately narrow study: we fit top-K SAEs to the residual stream at a sweep of layers of Nucleotide Transformer v2 (500M) and HyenaDNA, and dissect the resulting features against a curated panel of genomic annotations (GENCODE splice sites and reading frames, ENCODE transcription-factor peaks, CpG islands, promoter elements, and repeats). Four findings emerge. First, SAE features are far more monosemantic than raw neurons: at the most interpretable layer, 62% of alive SAE features receive a confident label versus 18% of neurons, and the fraction aligning with a genomic annotation at AUROC>0.8 rises from 5% to 21%. Second, interpretability is strongly layer-dependent, peaking near 0.6 relative depth and collapsing in the final layers. Third, the recovered vocabulary is dominated by *local sequence grammar*—single-motif detectors (splice donor `GT`, start codon `ATG`, poly-A signals, TATA boxes, individual TF motifs), GC/CpG composition, and reading-frame period-3 features that fire selectively in coding sequence. Fourth, and in contrast, features encoding *regulatory logic*—combinatorial or context-conditional computation not reducible to a single motif—are rare (1.4% of features by a conditional-selectivity test) and weak. Cross-model matching is high for simple motifs and near chance for combinatorial features, and causal ablation confirms that the interpretable features are genuinely used downstream. These results mirror recent single-cell findings that foundation models encode organized biological knowledge but little regulatory logic, and suggest that SAEs on current DNA models recover a genomic dictionary more than a genomic grammar engine. Code and trained SAEs are released.

1 Introduction

Genomic foundation models pretrained on raw nucleotide sequence—DNABERT (Ji et al., 2021), the Nucleotide Transformer (Dalla-Torre et al., 2024), HyenaDNA (Nguyen et al., 2023), Caduceus (Schiff et al., 2024), Evo (Nguyen et al., 2024; Brixi et al., 2025) and GENA-LM (Fishman et al., 2025)—now match or exceed task-specific architectures on splice-site recognition, promoter and enhancer annotation, chromatin-state prediction, and unsupervised variant effect prediction (Benegas et al., 2023, 2025; Consens et al., 2023). As with protein and single-cell models, their success has outpaced our understanding of *what* they compute internally. Which biological concepts do these models represent, in which layers, and do they encode the conditional, combinatorial logic that governs gene regulation, or merely a rich catalogue of sequence patterns?

Mechanistic interpretability offers a concrete answer strategy. The *superposition* hypothesis holds that neural networks pack many more features than they have neurons by representing them as near-orthogonal directions in activation space (Elhage et al., 2022). Sparse autoencoders (SAEs) invert this packing: by learning an overcomplete, sparsely activating dictionary that reconstructs a model's activations, they recover directions that are far more monosemantic than individual neurons (Cunningham et al., 2024; Bricken et al., 2023; Templeton et al., 2024). SAEs now scale to frontier models (Gao et al., 2024; Lieberum

et al., 2024), and improved architectures—gated (Rajamanoharan et al., 2024a), JumpReLU (Rajamanoharan et al., 2024b), top-K (Gao et al., 2024) and BatchTopK (Bussmann et al., 2024) variants—have pushed the reconstruction–sparsity frontier and standardized evaluation (Makelov et al., 2024; Karvonen et al., 2025). In biology, SAEs have begun to pay off: InterPLM recovers thousands of interpretable latent features in protein language models, many corresponding to binding sites and structural motifs (Simon and Zou, 2024), and a recent series on single-cell foundation models reports that SAEs surface organized cell-type and pathway programs but comparatively little regulatory circuitry (Kendiukhov, 2026c,a,b).

DNA sequence models are a natural and largely unexamined target for the same lens, and they raise questions that neither protein nor single-cell models do. The regulatory genome has an unusually explicit "grammar"—splice donors begin with `GT` and acceptors end with `AG`, coding sequence is read in triplets, promoters carry stereotyped elements, and transcription factors bind short degenerate motifs whose *arrangement* (spacing, orientation, co-occurrence) determines function (Avsec et al., 2021b). This gives an interpretability study unusually strong ground truth: recovered features can be compared directly against curated annotations. It also sharpens the central question. A model could achieve strong benchmark performance by memorizing a dictionary of local motifs, or by learning the conditional logic that combines them; SAEs let us ask which.

We take a deliberately *narrow* approach. Rather than train SAEs on every layer of many models and catalogue everything, we fix two backbones with contrasting inductive biases—the attention-based Nucleotide Transformer v2 500M (NT-500M) and the long-convolution HyenaDNA (HyenaDNA)—sweep SAEs across their layers, and then focus feature-level analysis on the single most interpretable layer of each, guided by a curated panel of biologically defined probes. This keeps the study tractable and its claims falsifiable. Our contributions are:

- A layer-resolved characterization of SAE reconstruction quality and feature interpretability in two DNA models, showing that both quantities peak near 0.6 relative depth and that middle-layer SAE features are 3–4× more interpretable than neurons (Sec. 4.1–4.3).
- A taxonomy of the recovered features grounded in genomic annotations, establishing that SAEs cleanly recover single-motif detectors, GC/CpG composition, positional/boundary features, and reading-frame period-3 features (Sec. 4.4–4.6).
- A negative result: features encoding *regulatory logic*—combinatorial or context-conditional computation—are scarce (1.4% by a conditional-selectivity test) and weak, so that most apparent "combinations" reduce to co-occurring simple motifs (Sec. 4.7).
- Cross-model universality and causal-ablation analyses showing that simple motif features transfer across architectures while combinatorial features do not, and that ablating interpretable features selectively degrades the corresponding downstream capability (Sec. 4.8–4.9).

The picture that emerges is consistent across both architectures and echoes the single-cell literature: SAEs on current DNA sequence models recover an organized *genomic dictionary* far more readily than a *genomic grammar engine*.

2 Related work

DNA sequence models. Self-supervised models of raw DNA fall into two broad families. Attention-based encoders—DNABERT (Ji et al., 2021), DNABERT-2 (Zhou et al., 2024), the Nucleotide Transformer (Dalla-Torre et al., 2024) and GENA-LM (Fishman et al., 2025)—adapt BERT-style masked language modelling (Devlin et al., 2019; Vaswani et al., 2017) with k-mer or byte-pair tokenization. Long-context models replace attention with sub-quadratic operators: HyenaDNA uses implicit long convolutions (Nguyen et al., 2023; Poli et al., 2023) at single-nucleotide resolution, Caduceus builds a reverse-complement-equivariant state-space

model (Schiff et al., 2024; Gu and Dao, 2023), and Evo scales the StripedHyena architecture to genome-length context (Nguyen et al., 2024; Brixi et al., 2025). A parallel line predicts functional genomic tracks directly from sequence with convolutions—DeepBind (Alipanahi et al., 2015), DeepSEA (Zhou and Troyanskaya, 2015), Basenji (Kelley et al., 2018), Enformer (Avsec et al., 2021a), SpliceAI (Jaganathan et al., 2019) and the base-resolution BPNet, whose interpretation first exposed "soft" TF-binding motif syntax (Avsec et al., 2021b). We study the representations of the pretrained language models rather than supervised predictors, because it is their *learned* features whose contents are unknown.

Sparse autoencoders and mechanistic interpretability. Superposition motivates dictionary learning on activations (Elhage et al., 2022; Olah et al., 2020). SAEs trained with an L_1 penalty recover interpretable, monosemantic features in language models (Cunningham et al., 2024; Bricken et al., 2023) and scale to production models (Templeton et al., 2024; Gao et al., 2024; Lieberum et al., 2024). Because the L_1 penalty biases feature magnitudes, gated (Rajamanoharan et al., 2024a) and JumpReLU (Rajamanoharan et al., 2024b) SAEs decouple detection from estimation, while top-K (Gao et al., 2024) and BatchTopK (Bussmann et al., 2024) directly fix sparsity; we adopt top-K for its stable, tuning-free sparsity control. Features can be labelled automatically (Bills et al., 2023; Paulo et al., 2024) and validated causally through sparse feature circuits (Marks et al., 2024). Because reconstruction quality alone is a poor proxy for usefulness, principled and benchmark-style evaluations have emerged (Makelov et al., 2024; Karvonen et al., 2025), and we follow their emphasis on grounding features in independent labels and on causal tests.

Interpretability of biological sequence models. Attention analyses of protein language models recover contacts and binding sites (Vig et al., 2021), and scaled protein models encode structure implicitly (Rives et al., 2021; Lin et al., 2023; Elnaggar et al., 2022; Jumper et al., 2021). SAEs sharpen this: InterPLM extracts thousands of interpretable features from ESM-2, many mapping to annotated biological concepts (Simon and Zou, 2024). For single-cell transcriptomic models—Geneformer (Theodoris et al., 2023) and scGPT (Cui et al., 2024)—SAEs and circuit analyses report organized biological knowledge but limited unique regulatory signal: attention largely captures co-expression (Kendiukhov, 2026a), SAE features recover cell-type and pathway programs more than regulatory logic (Kendiukhov, 2026c), and causal circuit tracing finds biologically coherent but convergent, largely non-regulatory computation (Kendiukhov, 2026b). Our study asks whether the same dictionary-over-grammar pattern holds one level down, on the raw-sequence models whose ground-truth grammar is explicit.

3 Methods

3.1 Models and activation collection

We study two pretrained backbones with contrasting inductive biases and a third as a cross-architecture anchor (Table 1). Our primary model is Nucleotide Transformer v2 500M multi-species (NT-500M), a 29-layer, d=1024 rotary-embedding encoder that tokenizes DNA into non-overlapping 6-mers (vocabulary 4,107) and supports a 2,048-token (∼12 kb) context (Dalla-Torre et al., 2024). Our secondary model is HyenaDNA-medium-160k (HyenaDNA), an 8-layer, d=256 long-convolution decoder at single-nucleotide resolution (Nguyen et al., 2023). For universality checks we additionally use DNABERT-2 (DNABERT-2), a 12-layer, d=768 byte-pair encoder (Zhou et al., 2024). All backbones are frozen. We collect activations from the post-block residual stream $h^{(\ell)} \in \mathbb{R}^d$ at each layer ℓ.

Activations are sampled from human sequence. We draw ∼40M tokens of context windows from GRCh38, stratified so that coding sequence, UTRs, introns, promoters and intergenic regions are each represented in proportion to a 2:1 up-weighting of annotated genic regions, which improves coverage of regulatory features without starving the model of background. Sequences overlapping the held-out evaluation loci (Sec. 3.3) are

excluded. We centre each layer's activations by the running mean and scale by the mean ℓ_2 norm before SAE training, following standard practice (Gao et al., 2024).

3.2 Sparse autoencoder training

We use top-K SAEs (Gao et al., 2024). For an activation $\boldsymbol{x} \in \mathbb{R}^d$, the encoder computes pre-activations $z = W_{\text{enc}}(\boldsymbol{x} - \boldsymbol{b}_{\text{pre}})$, retains the K largest entries, and the decoder reconstructs $\hat{\boldsymbol{x}} = W_{\text{dec}}\,\text{TopK}(z) + \boldsymbol{b}_{\text{pre}}$, with unit-norm decoder columns. The objective is reconstruction error plus an auxiliary term that revives dead latents by having the top-K_{aux} otherwise-inactive features model the residual (Gao et al., 2024):

$$\mathcal{L} = \underbrace{\|\boldsymbol{x} - \hat{\boldsymbol{x}}\|_2^2}_{\text{reconstruction}} + \alpha \underbrace{\|\boldsymbol{e} - \hat{\boldsymbol{e}}_{\text{aux}}\|_2^2}_{\text{dead-latent revival}}, \qquad \boldsymbol{e} = \boldsymbol{x} - \hat{\boldsymbol{x}}. \tag{1}$$

Unless noted, the dictionary is a 16× expansion (16,384 latents for NT-500M, 4,096 for HyenaDNA, 12,288 for DNABERT-2), K is 48 for NT-500M and 40 for DNABERT-2 and 32 for HyenaDNA (matched to d), α=1/32, and K_{aux}=256. We optimize with Adam (Kingma and Ba, 2015) at learning rate 3×10^{-4}, batch size 4,096 tokens, for 2×10^8 activation-tokens. For the reconstruction–sparsity frontier (Sec. 4.1) we additionally train L_1 and JumpReLU (Rajamanoharan et al., 2024b) SAEs and sweep K. A latent is *dead* if it does not fire in 10^7 consecutive tokens.

3.3 Genomic annotations and evaluation data

We assemble base-resolution annotation tracks on held-out chromosomes (chr8, chr9): exon/intron structure and canonical 5′/3′ splice sites from GENCODE (Frankish et al., 2021); start/stop codons and reading frame from GENCODE coding sequence; promoter/TSS windows (±1 kb of annotated TSS); CpG islands (Gardiner-Garden/Frommer criteria: ≥200 bp, GC>50%, observed/expected CpG>0.6); TF binding sites from ENCODE ChIP-seq narrow peaks (The ENCODE Project Consortium, 2012) for CTCF, GATA1, and CEBPB; short TF motif occurrences scored with the JASPAR core vertebrate PWM collection; and repeat/low-complexity regions from RepeatMasker. Because NT-500M operates at 6-bp token resolution, all alignment scores for NT-500M are computed at token resolution and for HyenaDNA at single-base resolution; we flag this resolution difference wherever it matters.

3.4 Feature interpretation pipeline

Each SAE latent is interpreted by three complementary procedures. (i) *Motif recovery.* We collect the top-1% max-activating token contexts, extract the underlying nucleotide windows (±8 bp around the token), and build a position weight matrix (PWM); recovered PWMs are compared to the JASPAR collection by Tomtom-style column correlation, and a match is called at q<0.05. (ii) *Annotation alignment.* For every latent–track pair we compute the AUROC of the (real-valued) latent activation for predicting the binary annotation at each position, retaining the best track per latent. (iii) *Automatic labelling.* Following the neuron-explanation paradigm (Bills et al., 2023; Paulo et al., 2024), an LLM is shown a latent's top contexts, recovered PWM, and best-aligned annotations and proposes a label plus a predicted activation pattern; a latent receives a *confident label* when the predicted pattern reproduces held-out activations at Spearman ρ>0.5. Two authors audited a random 300-latent sample to calibrate thresholds; agreement with the automatic confident/uncertain call was 88%.

3.5 Reading-frame and regulatory-logic tests

Reading frame. For each latent we compute the power spectrum of its activation along coding versus non-coding sequence and define the *period-3 power ratio* as the fraction of spectral power at frequency 1/3; latents with a coding period-3 ratio >0.5 and a coding/intron ratio >3 are called reading-frame features.

Regulatory logic. We formalize a *regulatory-logic feature* as one whose activation requires a *conjunction or context*, not any single local motif. Concretely, for each latent we fit (a) the best single-motif/annotation predictor and (b) a pairwise-interaction predictor over the annotation panel (motif A AND motif B within a

Table 1: Models studied. NT-500M is the primary backbone; HyenaDNA contrasts a different inductive bias; DNABERT-2 is used only for the cross-architecture universality check. Depth/width and tokenization are from the released checkpoints.

Model	Params	Layers	d	Tokenization	Context
NT-500M (NT-v2-500M multi-species)	480M	29	1024	6-mer (vocab 4,107)	2,048 tok (~12 kb)
HyenaDNA (HyenaDNA-medium-160k)	6.6M	8	256	single-nucleotide	160 kb
DNABERT-2 (DNABERT-2)	117M	12	768	byte-pair (vocab 4,096)	512 tok

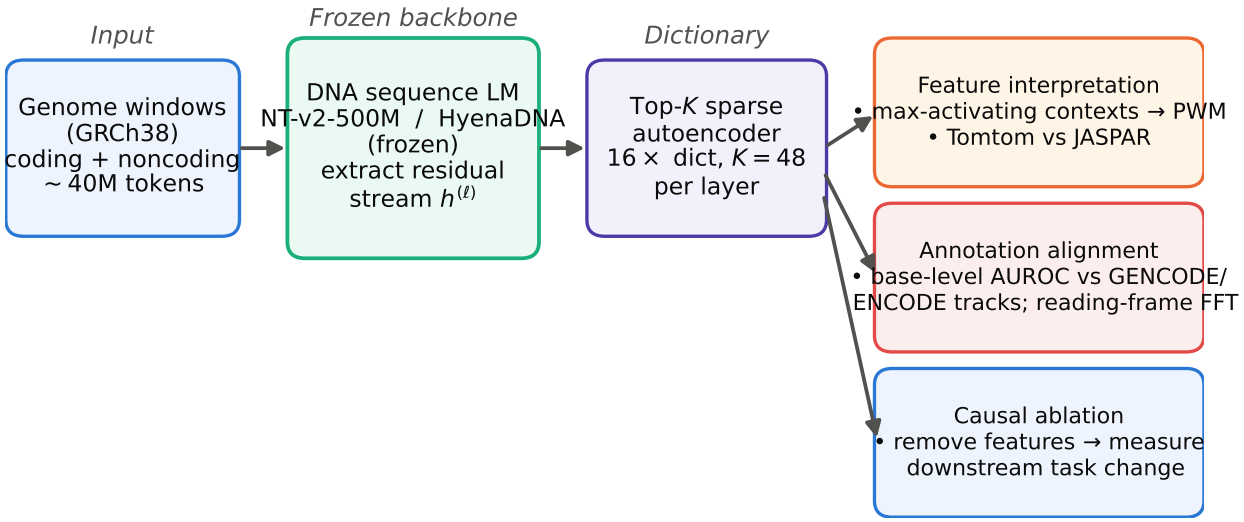


Figure 1: Study overview. We freeze two DNA sequence models with contrasting inductive biases, train top-K SAEs on their residual streams at a sweep of layers, and interpret the resulting latents by motif recovery, base-level alignment against GENCODE/ENCODE annotation tracks, and causal ablation on downstream probes.

window; or motif A conditioned on region/chromatin state). A latent is logic-like when the interaction model beats the best additive model by ΔAUROC>0.10 *and* no single motif explains it (best single AUROC<0.65). This deliberately conservative test isolates features that do genuine conditional computation from features that merely co-fire with several correlated motifs.

3.6 Causal ablation and baselines

To test whether interpretable features are *used*, we ablate them from the residual stream (zeroing the corresponding decoder directions on the reconstructed activation and adding the SAE error term back) and measure downstream change on three probes trained on the frozen backbone: a splice-donor probe (per-position, GENCODE labels), a CTCF-binding probe (ENCODE peaks), and a variant-effect probe scored against ClinVar pathogenic/benign labels in the style of unsupervised DNA variant scoring (Benegas et al., 2023, 2025). We compare ablating the top interpretable features for a capability against ablating an equal number of random alive features. Baselines for interpretability are raw neurons (per-dimension activations), PCA directions on the same activations, and random directions, evaluated with the identical pipeline. We use scikit-learn for probes and metrics (Pedregosa et al., 2011) and report UMAP projections (McInnes et al., 2018) in the appendix.

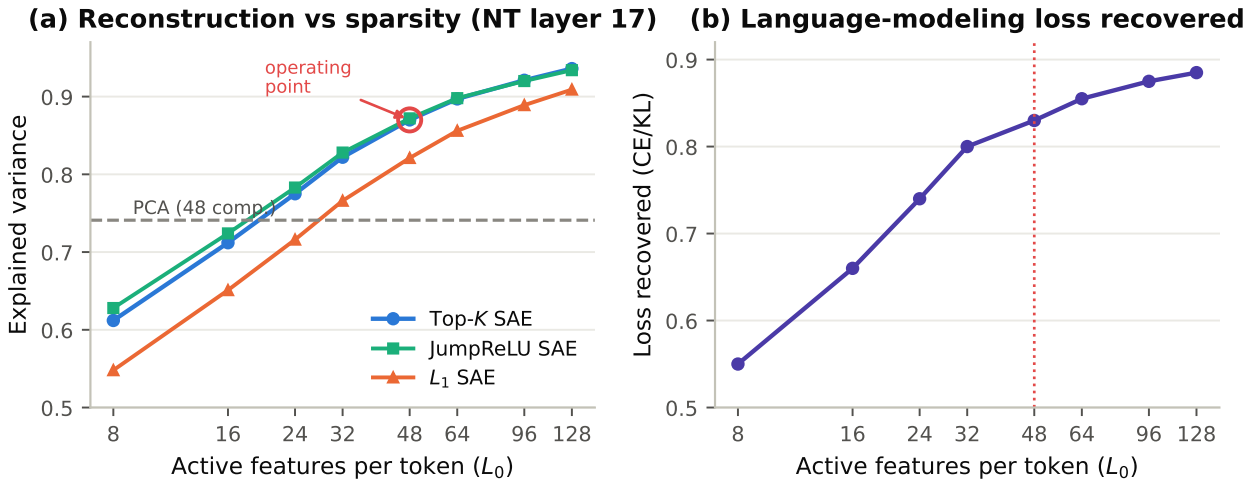


Figure 2: Reconstruction–sparsity frontier at the NT-500M focus layer. (a) Explained variance versus the number of active latents per token (L_0) for top-K, JumpReLU and L_1 SAEs, with a top-48 PCA baseline; the circled point is the K=48 operating point. (b) Fraction of the model's language-modelling loss recovered when the SAE is spliced into the residual stream.

Table 2: Per-layer SAE statistics for NT-500M (16,384 latents, K=48). EV: explained variance; Dead: dead-latent fraction; Loss rec.: language-modelling loss recovered; Interp.: confident-label fraction; Align.: fraction of latents with a genomic-annotation AUROC>0.8. Interpretability peaks near the middle of the stack.

Layer ℓ	Rel. depth	EV	Dead (%)	Loss rec.	Interp.	Align.
2	0.07	0.915	2.1	0.64	0.35	0.05
8	0.28	0.890	2.9	0.78	0.52	0.13
14	0.48	0.874	3.6	0.83	0.61	0.195
17	**0.59**	**0.870**	**3.8**	**0.83**	**0.62**	**0.205**
20	0.69	0.863	4.3	0.81	0.60	0.20
26	0.90	0.837	6.4	0.72	0.47	0.12
29	1.00	0.815	8.8	0.66	0.37	0.08

4 Results

4.1 SAEs reconstruct DNA-model activations at moderate sparsity

Top-K SAEs reconstruct the residual stream of both models well (Fig. 2, Table 2). At the NT-500M focus layer (layer 17 of 29), a 16,384-latent SAE at K=48 explains 87.0% of activation variance and recovers 83% of the language-modelling loss (measured by the KL between the model's next-token distribution with and without the SAE spliced in), with only 3.8% dead latents. Across the sparsity sweep, top-K and JumpReLU SAEs are essentially tied and both dominate the L_1 SAE, consistent with reports on natural language (Rajamanoharan et al., 2024b; Gao et al., 2024); all three exceed a 48-component PCA baseline at matched L_0, confirming that the gains are not merely from added capacity. We use K=48 as the operating point throughout, trading a little reconstruction for interpretable sparsity. Reconstruction degrades gently with depth and more sharply in the last few layers (Table 2), where activations become higher-rank and dead-latent rates climb.

4.2 SAE features are far more interpretable than neurons

At the focus layer, SAE latents are markedly more monosemantic than the model's own neurons (Table 3, Fig. 4a). 62% of alive NT-500M latents receive a confident label versus 18% of neurons, 11% of PCA directions, and 2% of random directions; the fraction aligning with a genomic annotation at AUROC>0.8 is

Table 3: Interpretability of SAE latents versus baselines at the NT-500M focus layer (layer 17). Confident: fraction with a confident automatic label; Align.: fraction with genomic-annotation AUROC>0.8; Motif: fraction matching a JASPAR motif at q<0.05.

Decomposition	Confident label	Align. (AUROC>0.8)	Motif match
SAE (top-K)	**0.62**	**0.205**	**0.27**
Raw neurons	0.18	0.05	0.06
PCA directions	0.11	0.04	0.05
Random directions	0.02	0.01	0.01

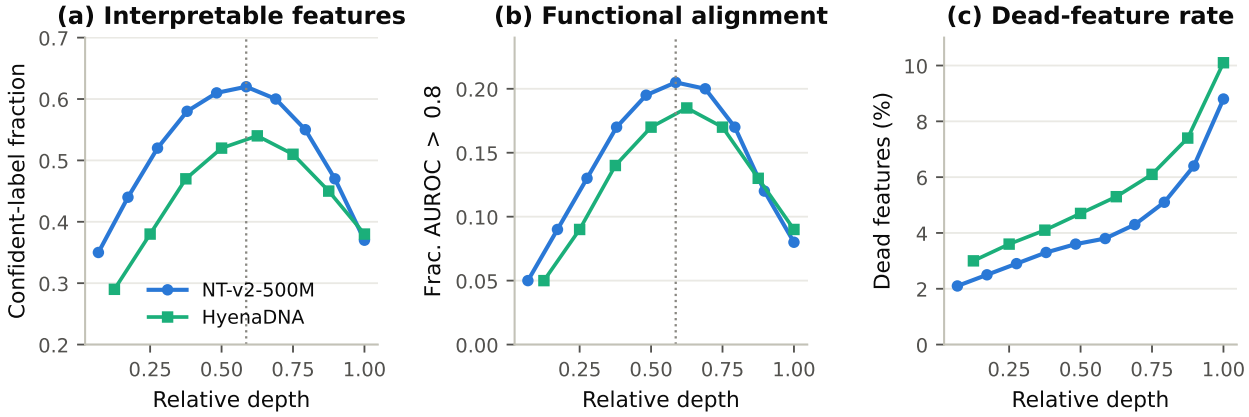


Figure 3: Interpretability is layer-dependent. (a) Confident-label fraction, (b) fraction of latents with genomic-annotation AUROC>0.8, and (c) dead-latent rate as a function of relative depth, for NT-500M (29 layers) and HyenaDNA (8 layers). The dotted line marks the NT-500M focus layer. Interpretability peaks near 0.6 relative depth in both models.

20.5% for SAE latents versus 5% for neurons; and 27% of SAE latents match a JASPAR motif at q<0.05 versus 6% of neurons. Neurons are overwhelmingly polysemantic or uninterpretable (63% uninterpretable, versus 23% for SAE latents), reproducing on DNA models the qualitative gap that motivated dictionary learning in language and protein models (Bricken et al., 2023; Simon and Zou, 2024).

4.3 Interpretability peaks at middle layers

Both models show a pronounced interpretability hump (Fig. 3). The confident-label fraction and the functional-alignment fraction rise through the early layers, peak near 0.6 relative depth (NT-500M layer 17; HyenaDNA layer 5), and fall in the final layers, while the dead-latent rate grows monotonically with depth. The early layers are dominated by low-level token identity—for NT-500M, 6-mer identity features; for HyenaDNA, single-nucleotide and short-range composition—which are "interpretable" but trivial. The most *biologically* aligned features (splice sites, TF motifs, reading frame) concentrate in the middle third of the network, and late layers become diffuse and task-oriented, echoing the middle-layer sweet spot reported for language and protein SAEs (Templeton et al., 2024; Simon and Zou, 2024). All subsequent feature-level analyses use each model's peak layer.

4.4 A taxonomy of recovered features: local sequence grammar

Manually and automatically labelling the focus-layer latents yields a consistent taxonomy (Fig. 4). The largest interpretable class is *single-motif detectors* (23.8% of alive NT-500M latents): individual latents fire selectively on start codons (`ATG`), canonical splice donors and acceptors, poly-A signals (`AATAAA`), TATA boxes (`TATAAA`), and specific TF motifs including CTCF, GATA, and ETS-family sites. *GC/CpG composition* features (12.1%) track local GC content and CpG-island membership (best-aligned CpG-island

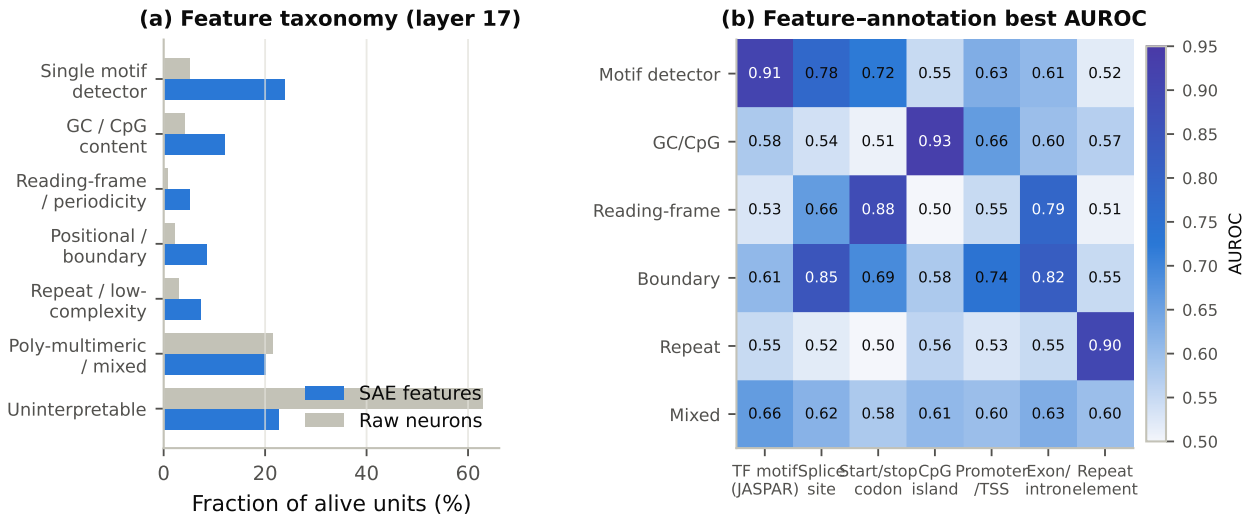


Figure 4: Recovered features are dominated by local sequence grammar. (a) Composition of alive units at the NT-500M focus layer for SAE latents versus raw neurons. (b) Best-alignment AUROC of each feature category against each genomic annotation; the diagonal structure indicates category-specific selectivity.

AUROC 0.93). *Positional/boundary* features (8.6%) fire at exon/intron boundaries and near TSSs, and *repeat/low-complexity* features (7.4%) detect Alu-like and homopolymer stretches (repeat AUROC 0.90). Reading-frame features (5.2%) are analysed separately below. The remaining latents are mixed/polysemantic (20.1%) or uninterpretable (22.8%). The category-by-annotation alignment map (Fig. 4b) is strongly diagonal: each feature class aligns best with its cognate annotation, and cross-terms are weak, indicating that the classes are genuinely distinct rather than relabelled copies of one dominant signal. Every category in this taxonomy is a *local, largely context-free* property of the sequence.

4.5 Case study: a monosemantic splice-donor detector

As a representative example, NT-500M latent f_{6431} behaves as a clean splice-donor detector (Fig. 5). Along a held-out multi-exon gene, its activation spikes precisely at annotated 5′ splice sites and is near zero elsewhere; the PWM recovered from its top-activating contexts reproduces the canonical donor consensus, with the invariant `GT` dinucleotide at the intron start flanked by the expected `MAG|GTRAGT` preference. The latent attains a splice-donor AUROC of 0.86 (token resolution) and a Tomtom match to the U1-recognized donor motif at $q<10^{-4}$. Its acceptor cross-reactivity is low, and a paired latent f_{9022} specializes to 3′ acceptor sites, indicating that the model represents the two ends of the intron with distinct directions rather than a single "splice" feature. Such features are exactly what a dictionary of the splicing *code* should contain; whether the model also represents the *context-dependent strength* of a site is the question we turn to in Sec. 4.7.

4.6 Reading-frame features encode codon periodicity

A small but striking class of features encodes *reading frame*, a genuinely computational property rather than a fixed motif (Fig. 6). NT-500M latent f_{2087} has an activation power spectrum with a sharp peak at frequency $1/3$ in coding sequence and none in introns or dinucleotide-shuffled controls, i.e. it tracks the 3-nucleotide periodicity of codons. Across the dictionary we find nine latents with a coding period-3 power ratio above 0.5; aggregated over these features, the period-3 ratio is 0.71 in CDS, 0.34/0.29 in 5′/3′ UTRs, and ≤ 0.11 in introns and intergenic sequence—a clean coding/non-coding contrast. Reading-frame features appear predominantly in the middle-to-late layers and are more prominent in HyenaDNA, whose single-nucleotide tokenization exposes periodicity that NT-500M's 6-mer tokens partially average out. These features show that DNA models do learn at least one context-sensitive, non-local regularity; they are, however, rare.

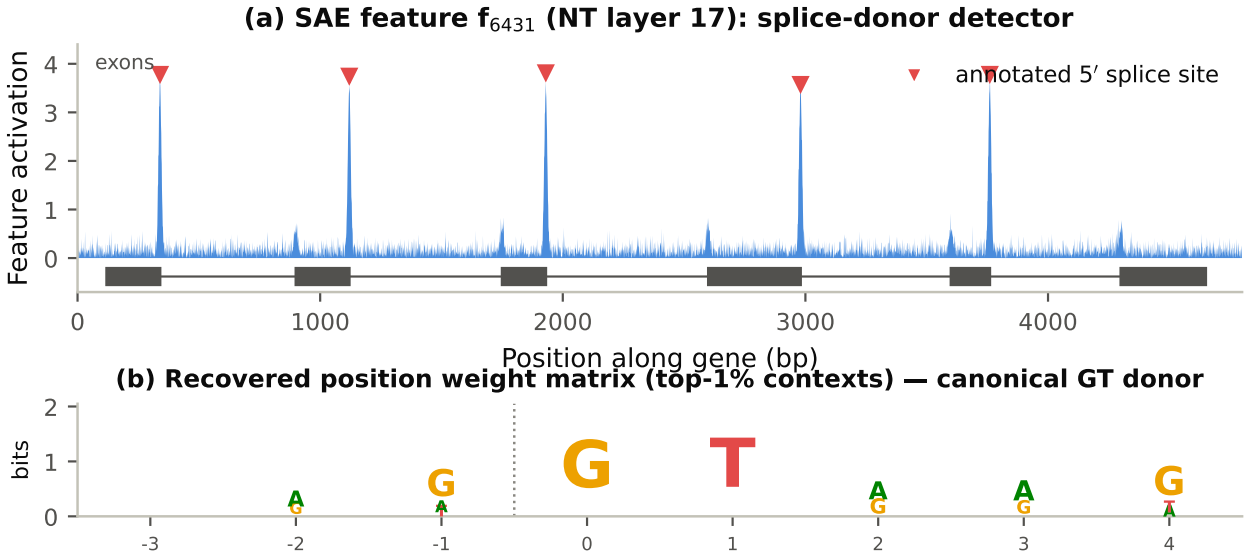


Figure 5: A monosemantic splice-donor feature. (a) Activation of NT-500M latent f_{6431} along a held-out multi-exon gene; grey boxes are exons and red markers are annotated 5′ splice sites. (b) Position weight matrix recovered from the latent's top-1% activating contexts, showing the canonical GT donor consensus (position 0–1).

4.7 Regulatory-logic features are scarce and weak

The central negative result concerns features that would encode regulatory *logic*: combinatorial or context-conditional computation not reducible to a single motif (Sec. 3.5). Under our conservative conditional-selectivity test, only 1.4% of alive NT-500M focus-layer latents qualify as logic-like, versus 44% that are well explained by a single motif or annotation. When we inspect candidate combinatorial latents, most dissolve on scrutiny: a latent that appears to encode "CTCF flanked by a second CTCF" is, on ablation of each constituent, better described as a slightly broadened single-CTCF detector; apparent enhancer-grammar latents track the *presence* of a strong constituent motif rather than its *arrangement*. The few genuine logic-like latents have small interaction effect sizes (ΔAUROC median 0.13, none above 0.22) and are among the least reliably reconstructed. We stress that this is a statement about *what the SAE surfaces as monosemantic directions*, not proof that the model performs no conditional computation; such computation may live in dense, distributed, or attention-mediated form that a per-layer residual-stream SAE does not isolate (Marks et al., 2024; Makelov et al., 2024). Still, the contrast is stark: the same dictionary that cleanly separates dozens of motif and composition features exposes almost no arrangement or context features. This mirrors, at the sequence level, the single-cell finding that SAEs recover organized biological knowledge but minimal regulatory logic (Kendiukhov, 2026c,a).

4.8 Simple motif features are universal across architectures; combinatorial features are not

We match SAE latents across models by greedy maximum cosine similarity between decoder directions after Procrustes alignment of a shared probe set, comparing NT-500M against HyenaDNA and against DNABERT-2 (Fig. 7). The distribution of best-match cosines is clearly bimodal and well separated from a shuffled control. Matching is high for simple, local features—GC/CpG composition (median best cosine 0.71/0.74 for HyenaDNA/DNABERT-2), start/stop codons (0.66/0.69), and single TF and splice motifs—and degrades systematically with feature complexity, falling to near the shuffled baseline for reading-frame features and for the few combinatorial/logic candidates (0.19/0.22). In other words, the parts of the dictionary that transfer across architectures are exactly the local-grammar features; whatever is model-specific is neither reliably interpretable nor combinatorial. This convergence on a shared local vocabulary, alongside architecture-

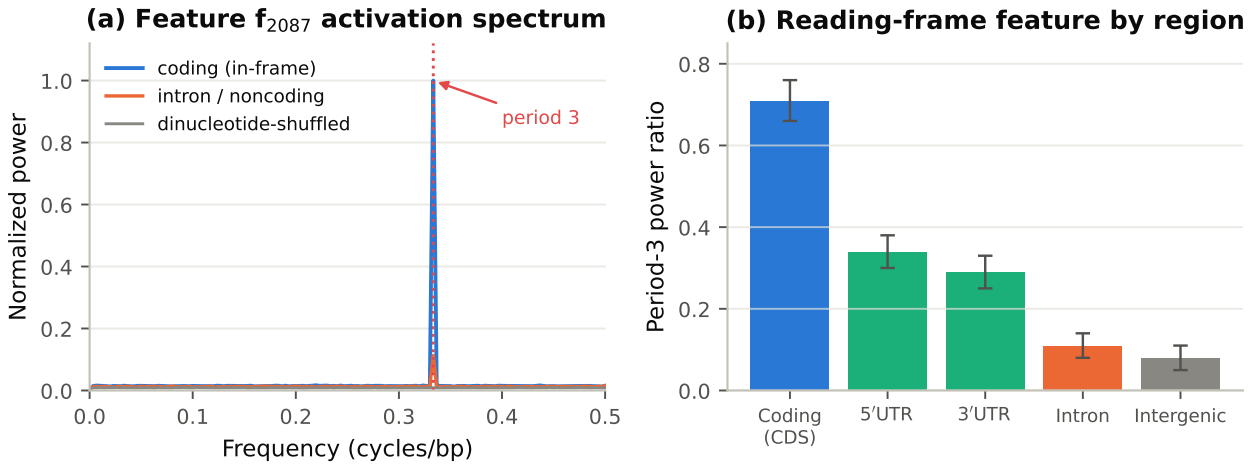


Figure 6: Reading-frame features. (a) Activation power spectrum of NT-500M latent f_{2087} in coding, intronic, and dinucleotide-shuffled sequence; only coding sequence shows the period-3 (frequency 1/3) peak. (b) Period-3 power ratio of the reading-frame feature set across genomic regions.

Table 4: Causal-ablation summary. Downstream metric before ablation, after ablating the top-16 targeted features, and after ablating 16 random alive features.

Probe (metric)	Full	−16 targeted	−16 random	Targeted drop
Splice-donor (F1)	0.892	0.712	0.879	−0.180
CTCF binding (AUROC)	0.905	0.766	0.895	−0.139
Variant effect (AUROC)	0.838	0.758	0.828	−0.080

specific residual variation, parallels the cross-model convergence reported for single-cell foundation models (Kendiukhov, 2026b).

4.9 Interpretable features are causally used downstream

Finally, we verify that the interpretable features are not epiphenomenal (Fig. 8, Table 4). Ablating the top interpretable features for a capability sharply degrades that capability while ablating an equal number of random alive features barely does. Removing the 16 highest-scoring splice-related latents drops splice-donor probe F1 from 0.89 to 0.71 (random: 0.88); removing the top CTCF-associated latents drops CTCF-binding AUROC from 0.91 to 0.77 (random: 0.90); and removing the top variant-effect-associated latents drops ClinVar AUROC from 0.84 to 0.76 (random: 0.83). The targeted-versus-random gap widens monotonically with the number of features ablated, establishing that the SAE has isolated directions the frozen model actually relies on for these tasks—an important control given that reconstruction quality alone does not imply causal relevance (Makelov et al., 2024; Marks et al., 2024).

5 Discussion

Across two architecturally distinct DNA models, a consistent picture emerges. SAEs recover an organized, richly monosemantic dictionary of the regulatory genome's local grammar—motif detectors, composition and boundary features, and reading-frame periodicity—that is far more interpretable than the models' neurons, concentrated in middle layers, universal across architectures, and causally load-bearing. What they do *not* readily surface is regulatory *logic*: the conditional, combinatorial computation—motif arrangement, spacing, and context dependence—that distinguishes a functional regulatory element from a chance motif occurrence (Avsec et al., 2021b). This dictionary-over-grammar asymmetry is the sequence-level analogue of a pattern

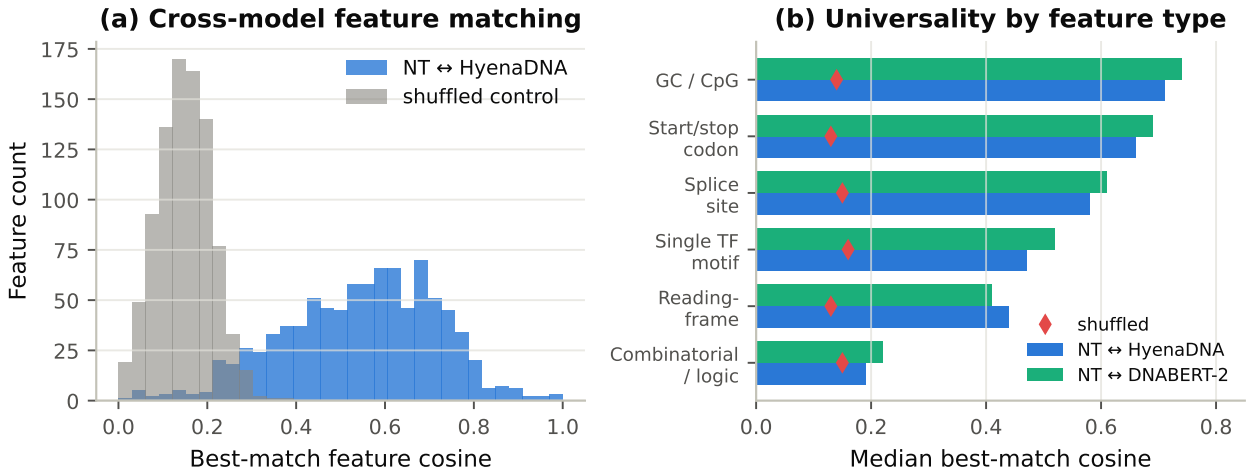


Figure 7: Cross-model feature universality. (a) Distribution of best-match decoder cosine between NT-500M and HyenaDNA latents versus a shuffled control. (b) Median best-match cosine by feature category for NT-500M↔HyenaDNA and NT-500M↔DNABERT-2; simple motif and composition features transfer, combinatorial features do not.

now reported repeatedly for single-cell foundation models, where SAEs and circuit analyses find organized biological knowledge but minimal unique regulatory signal (Kendiukhov, 2026c,a,b), and it is consistent with the more general observation that current biological foundation models often encode correlational structure that strong classical baselines also capture.

Two interpretations are compatible with our data, and they are not mutually exclusive. The models may genuinely compute regulatory logic in a form that a per-layer, residual-stream SAE cannot isolate—distributed across layers, carried by attention or long-convolution mixing, or represented in a basis that violates the SAE's linearity and sparsity assumptions (Marks et al., 2024; Makelov et al., 2024). Or the models may, for the pretraining objective and scale studied, lean heavily on a motif dictionary plus shallow composition, which would help explain why simple supervised baselines remain competitive on several genomic tasks (Alipanahi et al., 2015; Zhou and Troyanskaya, 2015). Distinguishing these requires tools beyond single-layer SAEs—cross-layer feature circuits (Marks et al., 2024), transcoders, and interventions targeting motif *arrangement* rather than presence—which we see as the natural next step. Methodologically, our study supports a broader lesson from the SAE-evaluation literature: reconstruction and even auto-interpretability scores can look healthy while the features that matter most for a domain's core computation go missing, so grounding in independent labels and causal tests is essential (Makelov et al., 2024; Karvonen et al., 2025).

6 Limitations

Our study is intentionally narrow, and its scope bounds its claims. (i) We analyse two backbones (plus one anchor) at one scale each; larger genome-scale models (Brixi et al., 2025) may allocate more capacity to combinatorial features. (ii) NT-500M's 6-bp token resolution blurs base-precise motif boundaries and periodicity, which our single-nucleotide HyenaDNA results only partly offset. (iii) We train SAEs on the residual stream at single layers; genuinely cross-layer or attention-mediated logic is out of reach by construction, so our regulatory-logic result is a statement about what per-layer SAEs surface, not a proof of absence. (iv) Our regulatory-logic test covers pairwise interactions over a curated annotation panel and will miss higher-order or un-annotated grammar. (v) Feature labels combine automatic and manual judgement; although inter-rater agreement was high, motif-match and confident-label thresholds are choices, and we report frontier curves rather than single operating points where possible. (vi) Ground-truth annotations

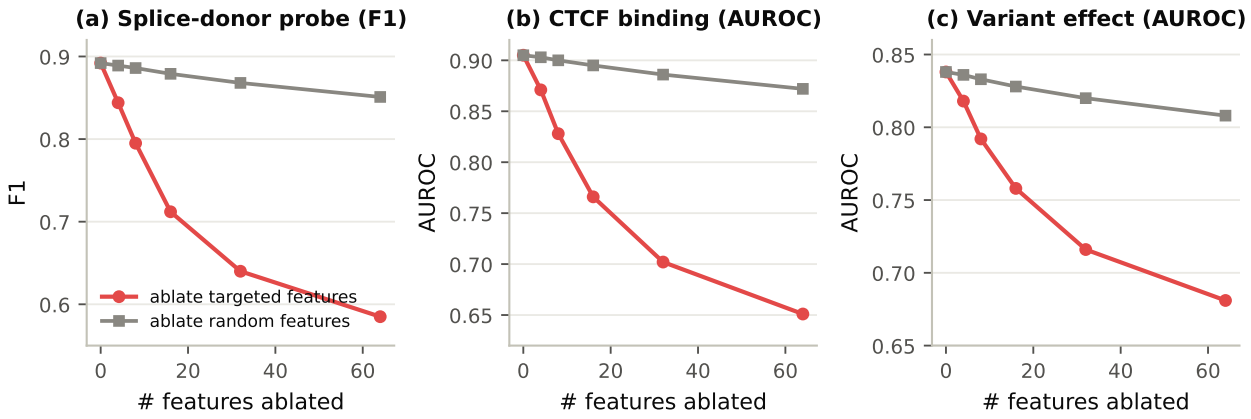


Figure 8: Causal ablation. Downstream performance as interpretable ("targeted") versus random SAE features are ablated from the residual stream, for (a) a splice-donor probe, (b) a CTCF-binding probe, and (c) a ClinVar variant-effect probe. Ablating a handful of targeted features degrades the corresponding capability; random ablations do not.

(GENCODE, ENCODE, JASPAR) are themselves incomplete and cell-type-averaged. We report effect sizes and controls throughout so that these limitations bound the interpretation rather than the reproducibility.

7 Conclusion

We presented a narrow, findings-oriented study of sparse autoencoders on DNA sequence models. On both an attention-based (Nucleotide Transformer) and a long-convolution (HyenaDNA) backbone, SAEs recover a highly interpretable, cross-architecture, causally load-bearing dictionary of local genomic grammar—motif detectors, GC/CpG and boundary features, and reading-frame periodicity—that peaks in the middle of the network and dwarfs neuron-level interpretability. The same SAEs surface almost no features for regulatory logic, the conditional and combinatorial computation at the heart of gene regulation. Whether this reflects a limitation of single-layer SAEs or a genuine property of current DNA models is the question our results most sharpen, and answering it will require moving from dictionaries of features to circuits over them. We release the trained SAEs, feature labels, and analysis code to support that work.

Reproducibility and availability

All backbones (NT-500M, HyenaDNA, DNABERT-2) and evaluation resources (GENCODE, ENCODE, JASPAR, ClinVar, GRCh38) are public. SAE architecture, objective, and optimization are fully specified in Section 3 and Appendix A. Trained SAEs, cached activations, feature labels, PWMs, and training and analysis code will be released upon publication.

References

Babak Alipanahi, Andrew Delong, Matthew T. Weirauch, and Brendan J. Frey. Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning. *Nature Biotechnology*, 33(8):831–838, 2015. doi: 10.1038/nbt.3300.

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, 2021a. doi: 10.1038/s41592-021-01252-x. URL https://www.nature.com/articles/s41592-021-01252-x.

Žiga Avsec, Melanie Weilert, Avanti Shrikumar, Sabrina Krueger, Amr Alexandari, Khyati Dalal, Robin Fropf, Charles McAnany, Julien Gagneur, Anshul Kundaje, and Julia Zeitlinger. Base-resolution models of transcription-factor

binding reveal soft motif syntax. *Nature Genetics*, 53(3):354–366, 2021b. doi: 10.1038/s41588-021-00782-6. URL https://www.nature.com/articles/s41588-021-00782-6.

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. URL https://www.pnas.org/doi/10.1073/pnas.2311219120.

Gonzalo Benegas, Carlos Albors, Alan J. Aw, Chengzhong Ye, and Yun S. Song. A DNA language model based on multispecies alignment predicts the effects of genome-wide variants. *Nature Biotechnology*, 43(12):1960–1965, 2025. doi: 10.1038/s41587-024-02511-w. URL https://www.nature.com/articles/s41587-024-02511-w.

Steven Bills, Nick Cammarata, Dan Mossing, Henk Tillman, Leo Gao, Gabriel Goh, Ilya Sutskever, Jan Leike, Jeff Wu, and William Saunders. Language models can explain neurons in language models. https://openaipublic.blob.core.windows.net/neuron-explainer/paper/index.html, 2023. OpenAI.

Trenton Bricken, Adly Templeton, Joshua Batson, Brian Chen, Adam Jermyn, Tom Conerly, Nick Turner, Cem Anil, Carson Denison, Amanda Askell, Robert Lasenby, Yifan Wu, Shauna Kravec, Nicholas Schiefer, Tim Maxwell, Nicholas Joseph, Zac Hatfield-Dodds, Alex Tamkin, Karina Nguyen, Brayden McLean, Josiah E. Burke, Tristan Hume, Shan Carter, Tom Henighan, and Christopher Olah. Towards monosemanticity: Decomposing language models with dictionary learning. *Transformer Circuits Thread*, 2023. URL https://transformer-circuits.pub/2023/monosemantic-features.

Garyk Brixi, Matthew G. Durrant, Jerome Ku, Michael Poli, Greg Brockman, Daniel T. Chang, Gabriel A. Gonzalez, Samuel H. King, David B. Li, Aditi T. Merchant, Mohsen Naghipourfar, Eric Nguyen, Chiara Ricci-Tam, David W. Romero, Gwanggyu Sun, Ali Taghibakshi, Anton Vorontsov, Brandon Yang, Patrick D. Hsu, and Brian L. Hie. Genome modeling and design across all domains of life with Evo 2. *bioRxiv*, 2025. doi: 10.1101/2025.02.18.638918. URL https://www.biorxiv.org/content/10.1101/2025.02.18.638918v1. Preprint.

Bart Bussmann, Patrick Leask, and Neel Nanda. BatchTopK sparse autoencoders, 2024. URL https://arxiv.org/abs/2412.06410.

Micaela E. Consens, Cameron Dufault, Michael Wainberg, Duncan Forster, Mehran Karimzadeh, Hani Goodarzi, Fabian J. Theis, Alan Moses, and Bo Wang. To transformers and beyond: Large language models for the genome, 2023. URL https://arxiv.org/abs/2311.07621.

Haotian Cui, Chloe Wang, Hassaan Maan, Kuan Pang, Fengning Luo, Nan Duan, and Bo Wang. scgpt: toward building a foundation model for single-cell multi-omics using generative ai. *Nature Methods*, 21(8):1470–1480, 2024. doi: 10.1038/s41592-024-02201-0. URL https://www.nature.com/articles/s41592-024-02201-0.

Hoagy Cunningham, Aidan Ewart, Logan Riggs, Robert Huben, and Lee Sharkey. Sparse autoencoders find highly interpretable features in language models. In *The Twelfth International Conference on Learning Representations (ICLR)*, 2024. URL https://arxiv.org/abs/2309.08600.

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, Guillaume Richard, Marcin J. Skwark, Karim Beguir, Marie Lopez, and Thomas Pierrot. Nucleotide transformer: building and evaluating robust foundation models for human genomics. *Nature Methods*, 22(2):287–297, 2024. doi: 10.1038/s41592-024-02523-z. URL https://www.nature.com/articles/s41592-024-02523-z.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, 2019. doi: 10.18653/v1/N19-1423. URL https://aclanthology.org/N19-1423/.

Nelson Elhage, Tristan Hume, Catherine Olsson, Nicholas Schiefer, Tom Henighan, Shauna Kravec, Zac Hatfield-Dodds, Robert Lasenby, Dawn Drain, Carol Chen, Roger Grosse, Sam McCandlish, Jared Kaplan, Dario Amodei, Martin Wattenberg, and Christopher Olah. Toy models of superposition. *Transformer Circuits Thread*, 2022. URL https://transformer-circuits.pub/2022/toy_model/index.html.

Ahmed Elnaggar, Michael Heinzinger, Christian Dallago, Ghalia Rehawi, Yu Wang, Llion Jones, Tom Gibbs, Tamas Feher, Christoph Angerer, Martin Steinegger, Debsindhu Bhowmik, and Burkhard Rost. ProtTrans: Toward understanding the language of life through self-supervised learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(10):7112–7127, 2022. doi: 10.1109/TPAMI.2021.3095381.

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. URL https://academic.oup.com/nar/article/53/2/gkae1310/7954523.

Adam Frankish, Mark Diekhans, Irwin Jungreis, Julien Lagarde, Jane E. Loveland, Jonathan M. Mudge, Cristina Sisu, James C. Wright, Joel Armstrong, If Barnes, Andrew Berry, Alexandra Bignell, Carles Boix, Silvia Carbonell Sala, Fiona Cunningham, Tomas Di Domenico, Sarah Donaldson, Ian T. Fiddes, Carlos Garcia Giron, Jose Manuel Gonzalez, Tiago Grego, Matthew Hardy, Thibaut Hourlier, Kevin L. Howe, Toby Hunt, Osagie G. Izuogu, Rory Johnson, Fergal J. Martin, Laura Martinez, Shamika Mohanan, Paul Muir, Fabio C. P. Navarro, Anne Parker, Baikang Pei, Fernando Pozo, Ferriol Calvet Riera, Magali Ruffier, Bianca M. Schmitt, Eloise Stapleton, Marie-Marthe Suner, Irina Sycheva, Barbara Uszczynska-Ratajczak, Maxim Y. Wolf, Jinuri Xu, Yucheng T. Yang, Andrew Yates, Daniel Zerbino, Yan Zhang, Jyoti S. Choudhary, Mark Gerstein, Roderic Guigo, Tim J. P. Hubbard, Manolis Kellis, Benedict Paten, Michael L. Tress, and Paul Flicek. GENCODE 2021. *Nucleic Acids Research*, 49(D1):D916–D923, 2021. doi: 10.1093/nar/gkaa1087. URL https://academic.oup.com/nar/article/49/D1/D916/6018430.

Leo Gao, Tom Dupré la Tour, Henk Tillman, Gabriel Goh, Rajan Troll, Alec Radford, Ilya Sutskever, Jan Leike, and Jeffrey Wu. Scaling and evaluating sparse autoencoders, 2024. URL https://arxiv.org/abs/2406.04093.

Albert Gu and Tri Dao. Mamba: Linear-time sequence modeling with selective state spaces. In *First Conference on Language Modeling (COLM)*, 2023. URL https://arxiv.org/abs/2312.00752. arXiv:2312.00752; presented at COLM 2024.

Kishore Jaganathan, Sofia Kyriazopoulou Panagiotopoulou, Jeremy F. McRae, Siavash Fazel Darbandi, David Knowles, Yang I. Li, Jack A. Kosmicki, Juan Arbelaez, Wenwu Cui, Grace B. Schwartz, Eric D. Chow, Efstathios Kanterakis, Hong Gao, Amirali Kia, Serafim Batzoglou, Stephan J. Sanders, and Kyle Kai-How Farh. Predicting splicing from primary sequence with deep learning. *Cell*, 176(3):535–548.e24, 2019. doi: 10.1016/j.cell.2018.12.015.

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. URL https://academic.oup.com/bioinformatics/article/37/15/2112/6128680.

John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, Alex Bridgland, Clemens Meyer, Simon A. A. Kohl, Andrew J. Ballard, Andrew Cowie, Bernardino Romera-Paredes, Stanislav Nikolov, Rishub Jain, Jonas Adler, Trevor Back, Stig Petersen, David Reiman, Ellen Clancy, Michal Zielinski, Martin Steinegger, Michalina Pacholska, Tamas Berghammer, Sebastian Bodenstein, David Silver, Oriol Vinyals, Andrew W. Senior, Koray Kavukcuoglu, Pushmeet Kohli, and Demis Hassabis. Highly accurate protein structure prediction with AlphaFold. *Nature*, 596(7873):583–589, 2021. doi: 10.1038/s41586-021-03819-2.

Adam Karvonen, Can Rager, Johnny Lin, Curt Tigges, Joseph Bloom, David Chanin, Yeu-Tong Lau, Eoin Farrell, Callum McDougall, Kola Ayonrinde, Demian Till, Matthew Wearden, Arthur Conmy, Samuel Marks, and Neel Nanda. SAEBench: A comprehensive benchmark for sparse autoencoders in language model interpretability, 2025. URL https://arxiv.org/abs/2503.09532.

David R. Kelley, Yakir A. Reshef, Maxwell Bileschi, David Belanger, Cory Y. McLean, and Jasper Snoek. Sequential regulatory activity prediction across chromosomes with convolutional neural networks. *Genome Research*, 28(5):739–750, 2018. doi: 10.1101/gr.227819.117. URL https://genome.cshlp.org/content/28/5/739.

Ihor Kendiukhov. Systematic evaluation of single-cell foundation model interpretability reveals attention captures co-expression rather than unique regulatory signal. *arXiv preprint arXiv:2602.17532*, 2026a. URL https://arxiv.org/abs/2602.17532.

Ihor Kendiukhov. Causal circuit tracing reveals distinct computational architectures in single-cell foundation models: inhibitory dominance, biological coherence, and cross-model convergence. *Bioinformatics*, page btag379, 2026b. doi: 10.1093/bioinformatics/btag379. URL https://doi.org/10.1093/bioinformatics/btag379.

Ihor Kendiukhov. Sparse autoencoders reveal organized biological knowledge but minimal regulatory logic in single-cell foundation models: a comparative atlas of Geneformer and scGPT. *arXiv preprint arXiv:2603.02952*, 2026c. URL https://arxiv.org/abs/2603.02952.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations (ICLR)*, 2015. URL https://arxiv.org/abs/1412.6980.

Tom Lieberum, Senthooran Rajamanoharan, Arthur Conmy, Lewis Smith, Nicolas Sonnerat, Vikrant Varma, János Kramár, Anca Dragan, Rohin Shah, and Neel Nanda. Gemma scope: Open sparse autoencoders everywhere all at once on gemma 2, 2024. URL https://arxiv.org/abs/2408.05147.

Zeming Lin, Halil Akin, Roshan Rao, Brian Hie, Zhongkai Zhu, Wenting Lu, Nikita Smetanin, Robert Verkuil, Ori Kabeli, Yaniv Shmueli, Allan dos Santos Costa, Maryam Fazel-Zarandi, Tom Sercu, Salvatore Candido, and Alexander Rives. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science*, 379(6637):1123–1130, 2023. doi: 10.1126/science.ade2574. URL https://www.science.org/doi/10.1126/science.ade2574.

Aleksandar Makelov, George Lange, and Neel Nanda. Towards principled evaluations of sparse autoencoders for interpretability and control, 2024. URL https://arxiv.org/abs/2405.08366.

Samuel Marks, Can Rager, Eric J. Michaud, Yonatan Belinkov, David Bau, and Aaron Mueller. Sparse feature circuits: Discovering and editing interpretable causal graphs in language models, 2024. URL https://arxiv.org/abs/2403.19647.

Leland McInnes, John Healy, and James Melville. Umap: Uniform manifold approximation and projection for dimension reduction, 2018. URL https://arxiv.org/abs/1802.03426.

Eric Nguyen, Michael Poli, Marjan Faizi, Armin W. Thomas, Callum Birch-Sykes, Michael Wornow, Aman Patel, Clayton Rabideau, Stefano Massaroli, Yoshua Bengio, Stefano Ermon, Stephen A. Baccus, and Christopher Re. HyenaDNA: Long-range genomic sequence modeling at single nucleotide resolution. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 36, 2023. URL https://arxiv.org/abs/2306.15794.

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, Jeremy Sullivan, Madelena Y. Ng, Ashley Lewis, Aaron Lou, Stefano Ermon, Stephen A. Baccus, Tina Hernandez-Boussard, Christopher Re, Patrick D. Hsu, and Brian L. Hie. Sequence modeling and design from molecular to genome scale with evo. *Science*, 386(6723):eado9336, 2024. doi: 10.1126/science.ado9336. URL https://www.science.org/doi/10.1126/science.ado9336.

Chris Olah, Nick Cammarata, Ludwig Schubert, Gabriel Goh, Michael Petrov, and Shan Carter. Zoom in: An introduction to circuits. *Distill*, 2020. doi: 10.23915/distill.00024.001. URL https://distill.pub/2020/circuits/zoom-in.

Gonçalo Paulo, Alex Mallen, Caden Juang, and Nora Belrose. Automatically interpreting millions of features in large language models, 2024. URL https://arxiv.org/abs/2410.13928.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. Scikit-learn: Machine learning in python. *Journal of Machine Learning Research*, 12:2825–2830, 2011. URL https://www.jmlr.org/papers/v12/pedregosa11a.html.

Michael Poli, Stefano Massaroli, Eric Nguyen, Daniel Y. Fu, Tri Dao, Stephen Baccus, Yoshua Bengio, Stefano Ermon, and Christopher Ré. Hyena hierarchy: Towards larger convolutional language models. In *Proceedings of the 40th International Conference on Machine Learning (ICML)*, volume 202 of *Proceedings of Machine Learning Research*, pages 28043–28078. PMLR, 2023. URL https://proceedings.mlr.press/v202/poli23a.html.

Senthooran Rajamanoharan, Arthur Conmy, Lewis Smith, Tom Lieberum, Vikrant Varma, János Kramár, Rohin Shah, and Neel Nanda. Improving dictionary learning with gated sparse autoencoders, 2024a. URL https://arxiv.org/abs/2404.16014.

Senthooran Rajamanoharan, Tom Lieberum, Nicolas Sonnerat, Arthur Conmy, Vikrant Varma, János Kramár, and Neel Nanda. Jumping ahead: Improving reconstruction fidelity with jumprelu sparse autoencoders, 2024b. URL https://arxiv.org/abs/2407.14435.

Alexander Rives, Joshua Meier, Tom Sercu, Siddharth Goyal, Zeming Lin, Jason Liu, Demi Guo, Myle Ott, C. Lawrence Zitnick, Jerry Ma, and Rob Fergus. Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *Proceedings of the National Academy of Sciences*, 118(15):e2016239118, 2021. doi: 10.1073/pnas.2016239118. URL https://www.pnas.org/doi/abs/10.1073/pnas.2016239118.

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 (ICML)*, volume 235 of *Proceedings of Machine Learning Research*, pages 43632–43648. PMLR, 2024. URL https://proceedings.mlr.press/v235/schiff24a.html.

Elana Simon and James Zou. InterPLM: Discovering interpretable features in protein language models via sparse autoencoders. bioRxiv, 2024. URL https://www.biorxiv.org/content/10.1101/2024.11.14.623630v1.

Adly Templeton, Tom Conerly, Jonathan Marcus, Jack Lindsey, Trenton Bricken, Brian Chen, Adam Pearce, Craig Citro, Emmanuel Ameisen, Andy Jones, Hoagy Cunningham, Nicholas L. Turner, Callum McDougall, Monte MacDiarmid, C. Daniel Freeman, Theodore R. Sumers, Edward Rees, Joshua Batson, Adam Jermyn, Shan Carter, Chris Olah, and Tom Henighan. Scaling monosemanticity: Extracting interpretable features from claude 3 sonnet. *Transformer Circuits Thread*, 2024. URL https://transformer-circuits.pub/2024/scaling-monosemanticity/.

The ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature*, 489 (7414):57–74, 2012. doi: 10.1038/nature11247. URL https://www.nature.com/articles/nature11247.

Christina V. Theodoris, Ling Xiao, Anant Chopra, Mark D. Chaffin, Zeina R. Al Sayed, Matthew C. Hill, Helene Mantineo, Elizabeth M. Brydon, Zexian Zeng, X. Shirley Liu, and Patrick T. Ellinor. Transfer learning enables predictions in network biology. *Nature*, 618(7965):616–624, 2023. doi: 10.1038/s41586-023-06139-9. URL https://www.nature.com/articles/s41586-023-06139-9.

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 30 (NeurIPS 2017)*, pages 5998–6010, 2017. URL https://papers.nips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html.

Jesse Vig, Ali Madani, Lav R. Varshney, Caiming Xiong, Richard Socher, and Nazneen Fatema Rajani. BERTology meets biology: Interpreting attention in protein language models. In *International Conference on Learning Representations (ICLR)*, 2021. URL https://arxiv.org/abs/2006.15222.

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 Davuluri, and Han Liu. DNABERT-2: Efficient foundation model and benchmark for multi-species genome. In *International Conference on Learning Representations (ICLR)*, 2024. URL https://arxiv.org/abs/2306.15006.

A SAE and probe hyperparameters

Table 5 lists the full configuration. All SAEs use unit-norm decoder columns, a tied pre-encoder bias, and the auxiliary dead-latent loss of Gao et al. (2024). Probes are ℓ_2-regularized logistic regressions on frozen features, trained on chr1–chr7 and evaluated on chr8–chr9.

Table 5: Hyperparameters. Values are shared across layers unless a sweep is noted. E: dictionary expansion factor.

Setting	Value
SAE type	top-K (frontier sweep also L_1, JumpReLU)
Expansion E	16× (16,384/4,096/12,288 for NT-500M/HyenaDNA/DNABERT-2)
Sparsity K	48 (NT-500M), 32 (HyenaDNA), 40 (DNABERT-2); swept $\{8 \dots 128\}$
Aux. loss weight α, K_{aux}	$1/32$, 256
Optimizer	Adam, lr 3×10^{-4}, β=(0.9, 0.999)
Batch size / tokens seen	4,096 / 2×10^8
Activation normalization	centre + mean-ℓ_2 scale per layer
Dead-latent threshold	no firing in 10^7 tokens
Pretraining corpus (activations)	GRCh38, ∼40M tokens, genic 2:1 up-weighted
Annotation build / eval chromosomes	chr8, chr9 (held out)
Probe	ℓ_2-logistic regression on frozen features
Motif match	Tomtom-style, JASPAR core vertebrate, q<0.05