# CENO: A Genome-Scale World Model for Evolutionary Sequence Interpretation and Programmable Regulatory Design

**Mingqian Ma**[*1], **Yucheng Wu**[*1,2], **Xin Chen**[*1,3], **Feifei Jiang**[*1,4], **Peijun Lin**[5], **Dongxin Ye**[6,7], **Yidi Sun**[8], **Yijing Zhang**[2], **Tianqiong Shi**[9], **Yu Zhao**[10], **Wanli Ouyang**[11,12,1], **Bowen Zhou**[1,13], **Lei Bai**[*†1], **Yuchen Ren**[†1]

[1]Shanghai Artificial Intelligence Laboratory, [2]Fudan University, [3]The University of Sydney, [4]Shanghai Jiao Tong University, [5]University of Southern California, [6]Shanghai Innovation Institute, [7]University of Electronic Science and Technology of China, [8]Center for Excellence in Brain Science and Intelligence Technology, Chinese Academy of Sciences, [9]Nanjing Normal University, [10]Westlake University, [11]Shenzhen Loop Area Institute, [12]The Chinese University of Hong Kong, [13]Tsinghua University

## Abstract

DNA encodes biological function across a continuum of sequence scales, from single-nucleotide and motif-level grammar to regulatory neighborhoods, chromatin-scale organization and evolutionary constraint. A useful model of genomes should therefore do more than classify short sequence windows: it should maintain nucleotide-resolution state over long contexts, score counterfactual mutations, condition on homologous sequence evidence and generate sequences under functional constraints. We define such a system operationally as a genomic world model: a general-purpose generative model of genome sequence space that unifies sequence understanding and sequence design through a shared state and likelihood interface. Here we introduce CENO, a family of long-context generative genomic world models designed to preserve local DNA grammar while extending usable context to regulatory and chromatin scales. CENO combines Mamba sequence-mixing layers, sparse attention layers and mixture-of-experts capacity in a single autoregressive backbone, and is trained at 300M, 600M and 1B parameter scales with a staged curriculum that progresses from 8k-token cross-domain genomic pretraining to 131k- and 1M-token whole-genome long-context continuation. We evaluate CENO under a unified world-model benchmark paradigm spanning retrieval, representation, counterfactual perturbation, reconstruction, evolutionary conditioning and design. CENO retains practical long-context inference and retrieves distal sequence in synthetic assays. In zero-shot long-context analyses, without task-specific finetuning, long-context continuation yields annotation- and chromatin-boundary-associated attention patterns and frozen-state representations that generalize across human cell types and mouse cell or tissue settings. To incorporate evolutionary information, we further post-train CENO on packed real multiple-sequence-alignment contexts and score variants by reference–mutant likelihood deltas, improving matched variant-effect prediction and producing strong cross-species evolutionary-enrichment signals. Complementing these perturbation-based variant tests, we evaluate zero-shot long-sequence generation by partial-gene continuation, asking whether the model can recover withheld gene-scale sequence structure across eukaryotic, bacterial and archaeal species; recovery improves with model scale and later whole-genome long-context training. Finally, we use CENO as the backbone for a cell-type-specific enhancer design workflow in mouse cortex, coupling a CENO-based accessibility oracle with conditional supervised fine-tuning and oracle-guided reinforcement learning. Together, CENO provides a genome-scale world-model framework for sequence interpretation, evolutionary reasoning, gene-scale reconstruction and programmable regulatory sequence generation.

[*] Core Author [†] Corresponding Author.
Please send correspondence regarding this report to renyuchen@pjlab.org.cn

## 1. Introduction

Genome modeling is a short-to-long sequence-learning problem. At short ranges, DNA encodes nucleotide-resolution syntax, transcription-factor motifs, splice signals, codon structure and local regulatory grammar over tens to hundreds of bases. At longer ranges, the same sequence is organized into promoters, enhancers, repeats, genes, regulatory neighborhoods, chromatin domains and locus-scale structures that can span tens of kilobases to megabases. A useful genomic model should therefore not trade local precision for long context, or long context for local grammar. It should connect these scales in a single sequence model, learning how short sequence features are composed, routed and constrained across progressively longer genomic contexts.

We use the term genomic world model in an operational sense. The "world" is not a complete organismal or cellular simulator, but the sequence-level world in which local nucleotide grammar, gene structure, regulatory neighborhoods, chromatin-scale organization, evolutionary constraint and designable function jointly determine the plausibility and consequence of DNA sequences. Under this definition, a genomic world model should satisfy four requirements. First, it should be general-purpose, spanning coding, noncoding, regulatory and cross-species sequence regimes rather than a single supervised task. Second, it should be stateful over long contexts, so that distal sequence can influence local predictions and internal representations reflect locus-scale organization. Third, it should unify understanding and generation: the same model should score mutations, recover withheld sequence, condition on homologous context and generate new sequences. Fourth, it should be evaluated by a unified benchmark suite that probes retrieval, representation, perturbation, reconstruction, evolution and design rather than by a single leaderboard. In this sense, a genomic foundation model is defined by reuse, whereas a genomic world model is defined by reusable operations over genome sequence space: completion, perturbation, conditioning and design through a shared generative interface.

Recent DNA language models have advanced different parts of this goal. Encoder-style models such as DNABERT2, Nucleotide Transformer and Genomics-FM showed that large-scale pretraining can produce transferable genomic representations [1, 2, 3]. Long-sequence architectures such as HyenaDNA and Caduceus extended genomic modeling beyond the context lengths of standard Transformers [4, 5]. Generative models including Evo, Evo2, HybriDNA and Omnireg-gpt further moved the field toward autoregressive genome-scale sequence modeling and sequence generation [6, 7, 8, 9]. In parallel, supervised sequence-to-function models such as Enformer, Borzoi and AlphaGenome have shown that DNA sequence models can predict regulatory activity, expression and variant effects when trained directly on molecular phenotypes [10, 11, 12]. Together, these studies show that DNA models can learn local sequence grammar, long-range dependencies, molecular phenotypes and generative sequence distributions, but these capabilities are often developed and evaluated in separate modeling regimes.

This separation leaves an important gap for genomic world modeling. Long context is not equivalent to a large maximum input length. A model may accept long sequences while still failing to retrieve distant information, generate efficiently after long prompts, or form internal representations that reflect long-range genomic organization. Conversely, a model optimized for local or short-window tasks may perform well on motif, splicing or coding readouts while remaining insensitive to regulatory context beyond the immediate neighborhood. A practical long-context genomic world model should therefore satisfy three conditions: it should remain computationally usable at long input lengths; it should preserve short-range sequence grammar while routing information across distant loci; and it should expose representations that can be interrogated for biologically meaningful long-range structure.

Benchmark behavior also reflects this short-to-long tension. Performance on short regulatory classification, splice-altering variants, coding fitness, structural variants, long-context retrieval, chromatin-scale

diagnostics and sequence generation can favor different model families. Variant-effect prediction and long-sequence generation probe complementary aspects of genomic understanding. Variant-effect prediction asks whether a model assigns meaningful score changes to local sequence perturbations of an observed reference. Long-sequence generation instead asks whether a model can continue a partially observed gene or genomic sequence and recover the withheld gene-scale structure. The former tests perturbation sensitivity; the latter tests reconstruction of species-specific sequence organization, coding and noncoding grammar, and gene-level context. A genomic world model should therefore be evaluated not only by variant scores, but also by whether its generative distribution improves with scale and whole-genome long-context training.

Generalization across biological contexts is another requirement. A model that detects a chromatin-boundary-associated signal in one locus or one cell type may not have learned a reusable long-range representation. Similarly, a model that recovers sequence continuations in one species may not have learned a broadly useful genomic distribution. Long-context evaluation should therefore test whether model-internal signals transfer across cell types, tissues and species, and whether sequence recovery holds across diverse genomes rather than only within a single reference context. This generalization requirement is central to the world-model view: the learned sequence distribution should support operations that transfer across biological contexts, not only memorized local patterns.

Variant interpretation adds a further axis to this problem. Many variant effects are shaped not only by the target sequence itself, but also by evolutionary constraint across homologous sequences. Alignment- and phylogeny-aware models such as GPN-MSA and GPN-Star use multispecies genome alignments to improve variant-effect prediction [13, 14]. PoET, developed for protein families, further shows that family-level evolutionary context can provide a strong training signal for autoregressive sequence models [15]. These studies point to a missing capability for genomic world models: combining single-sequence genome pretraining with post-training that exposes the model to homologous context, while retaining a simple sequence-likelihood interface for downstream variant scoring.

A further goal is to move from interpretation to design. Genomic foundation models are useful not only when they score existing variants or annotate existing loci, but also when they can generate candidate sequences under functional constraints. Regulatory design makes the short-to-long problem especially clear. A designed enhancer must contain local motif grammar, preserve plausible sequence composition, and satisfy cell-type-specific functional constraints defined by a broader regulatory program. This requires a generative backbone that can be adapted with task-specific feedback without discarding the sequence grammar and contextual representations learned during pretraining. In the world-model framing, design is not a separate downstream add-on: it is one of the core operations that tests whether the learned sequence distribution can be steered toward new functional regions of genome space.

Here we introduce CENO, a family of long-context generative genomic world models designed around this short-to-long principle. CENO combines Mamba sequence-mixing layers, sparse attention layers and mixture-of-experts layers in a single causal autoregressive backbone. We train 300M, 600M and 1B scale models with a staged curriculum. The first stages learn broad genomic sequence grammar at 8k context across prokaryotic, metagenomic, viral, organelle, eukaryotic, transcript, splicing and regulatory sources. The later stages shift toward eukaryotic, transcript, splicing and regulatory sequence and then extend the context to 131k and 1M tokens using long windows from complete eukaryotic genomes. This curriculum separates three axes that are often entangled in DNA language modeling: model capacity, sequence length and data composition. It also creates a natural test of whether later whole-genome continuation improves gene-scale sequence recovery, rather than only increasing the configured context length.

CENO is also designed for evolutionary post-training. The pretrained model first learns single-sequence

genome syntax and long-range state tracking. We then adapt the same backbone to packed real multiple-sequence-alignment contexts, in which a target sequence is paired with homologous rows and explicit row identities. Early row-local layers preserve within-sequence modeling, whereas later fusion layers allow homologous rows to condition the target representation. At inference, variant effects are computed as likelihood differences between mutant and reference target sequences under the same MSA context. Thus, evolutionary information shapes the model during post-training and scoring, while the downstream readout remains a direct sequence-likelihood delta rather than a supervised task-specific classifier.

We evaluate CENO across the capabilities required of a genomic world model. We first measure pretraining scaling, long-prefill throughput and sustained generation speed, and test whether distant sequence influences terminal predictions in a synthetic DNA needle-in-a-haystack assay. We then probe whether long-context continuation induces annotation- and chromatin-boundary-associated signals, using zero-shot TAD-scale attention diagnostics, annotation-linked attention examples and frozen-state linear probes across human cell types and mouse cell or tissue settings. We next evaluate zero-shot genomic prediction across variant-effect, fitness, regulatory-impact and structural-variant benchmarks. Complementing these perturbation-based tests, we evaluate long-sequence generation by partial-gene continuation across eukaryotic, bacterial and archaeal species, treating withheld-sequence recovery as a reconstruction-style benchmark for gene-scale sequence understanding. We then evaluate MSA-context post-training for evolutionary variant scoring across matched BRCA1/BRCA2 comparisons, human disease and finemapping benchmarks, and cross-species evolutionary-enrichment tasks. Finally, we use CENO as the backbone for a cell-type-specific enhancer design workflow in mouse cortex, coupling a CENO-based accessibility oracle with conditional supervised fine-tuning and GRPO-based reinforcement learning. Together, these analyses position CENO as a genome-scale world-model framework that links local sequence grammar, long-range genome organization, counterfactual variant scoring, evolutionary conditioning, gene-scale sequence recovery and regulatory sequence generation.

## 2. Results

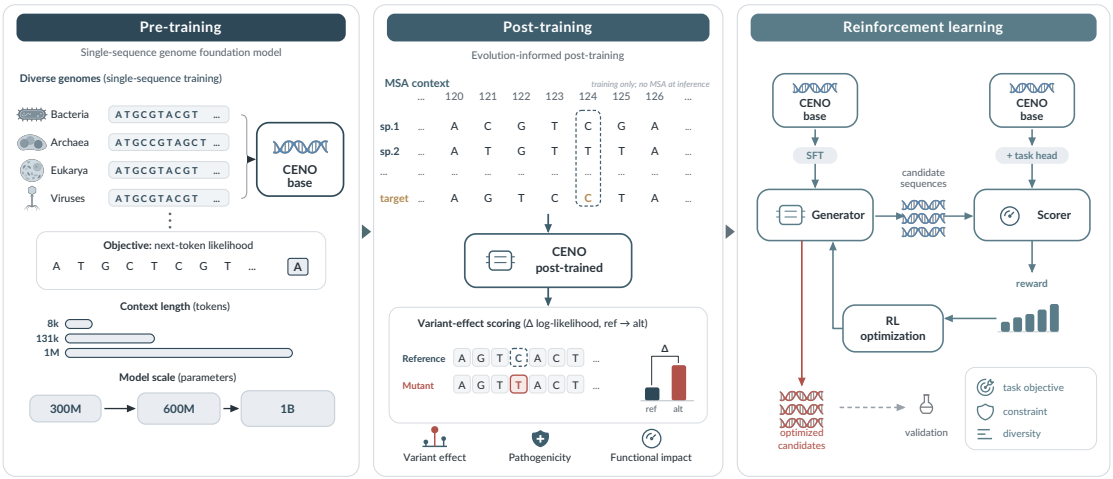
### 2.1. CENO pretraining spans local genome grammar to megabase context

CENO was trained as a causal genome-model family at approximately 300M, 600M and 1B total parameters. Figure 1a outlines a long-context pretraining program that links the model backbone, context curriculum and downstream interfaces. This design provides a shared foundation for using extended genomic context in analysis, downstream applications and sequence generation.

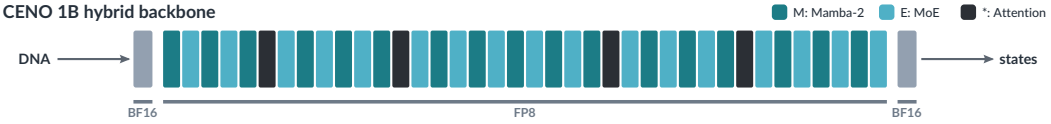
The backbone combines Mamba-2 sequence-mixing blocks, attention blocks and mixture-of-experts (MoE) capacity within a single causal architecture (Figure 1b, Methods 3.1). Across the 300M, 600M and 1B total parameter scales, the corresponding active parameter counts were approximately 100M, 200M and 400M per token.

The context curriculum trained Stage I/II models at 8,192 tokens. Stage III continued from the 8k checkpoint at 131,072 tokens. Stage IV continued from the 131k checkpoint at 1,048,576 tokens (Figure 1c). Stage I/II production recipes were selected by a constrained RegMix workflow (Methods 3.2). Candidate mixtures were sampled under eukaryotic/prokaryotic balance constraints, fitted from small-scale mixture-response validation experiments and selected with prespecified benchmark-response filters. The fixed production weights are reported in Supplementary Table S2 and source data. The additional selection archives needed to recreate the mixture search are identified separately. At the high-level category scale, this produced a near-balanced Stage I recipe: 51% prokaryotic, metagenomic, viral and organelle data, and 49% eukaryotic, transcript, splicing and regulatory data.

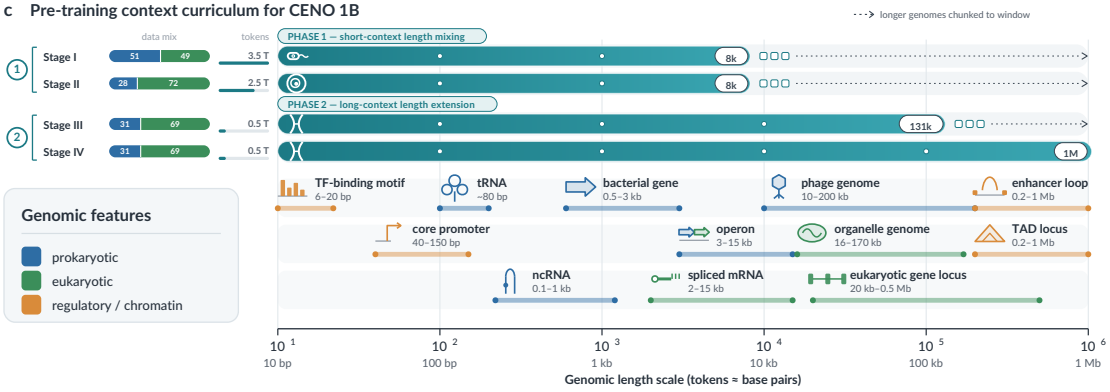
**a** Training-to-design paradigm



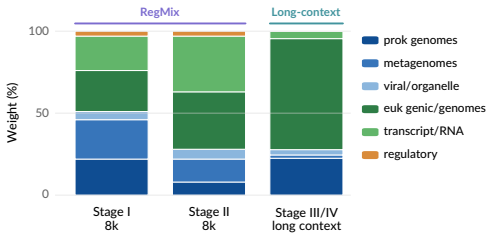

**b** CENO 1B hybrid backbone



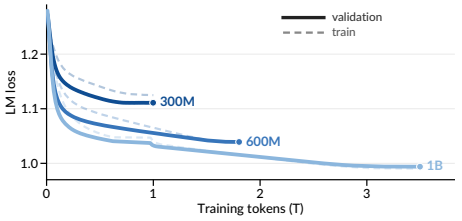

**c** Pre-training context curriculum for CENO 1B




**d** Dataset composition during training




**e** Stage I loss scaling




**f** Long-context inference throughput

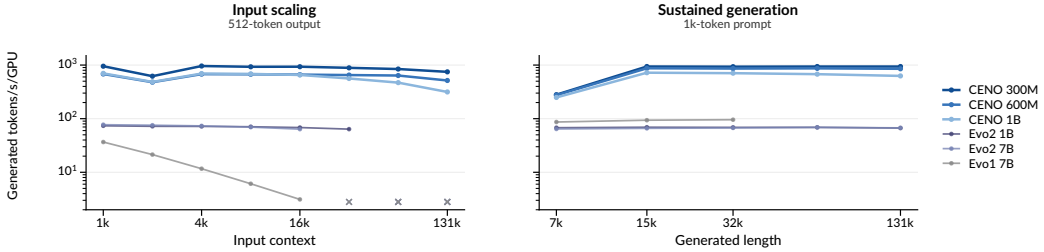


Figure 1 | **Overview of the CENO genomic foundation model. a,** Training-to-design paradigm. *Pre-training*: a single-sequence next-token-likelihood objective over diverse genomes (bacteria, archaea, eukarya, viruses) yields the CENO base model under a context-length curriculum (8k → 131k → 1M tokens) at three scales (300M/600M/1B). *Post-training*: real multiple-sequence alignments (MSA) supply evolutionary constraint; the post-trained model gives a variant-effect readout (VEP score) from the log-likelihood difference ($\Delta \log p$) between reference and alternate alleles, supporting VEP, pathogenicity and functional-impact tasks. *Reinforcement learning*: from the CENO base, supervised fine-tuning (SFT) and a task head yield a CENO generator and scorer; candidate sequences are generated and scored, and the reward drives RL optimization toward optimized candidates under task-objective, constraint and diversity goals (dashed arrow, validation). Colour encodes input/data, model, evaluation and output roles. **b,** CENO 1B hybrid backbone: a causal trunk interleaves Mamba-2 (M), mixture-of-experts (E) and attention (∗) blocks between input/output embeddings; embeddings in BF16, trunk compute in FP8. **c,** Biological-context length curriculum on a shared log length axis (tokens ≈ base pairs). Top: every stage trains on the full length range, with a solid bar for the contiguous context window and a segmented bar for longer genomes chunked to fit it (Stage I–II, 8k; Stage III, 131k; Stage IV, 1M). Bottom: representative genomic features at their characteristic length. **d,** Dataset composition during training: stacked weight (%) per stage; Stage I/II set by constrained RegMix, Stage III/IV by long-context weighting. **e,** Stage I loss scaling: validation (solid) and train (dashed) LM loss versus training tokens for 300M/600M/1B; the 1B model reaches the lowest endpoint. **f,** Long-context inference throughput (tokens $s^{-1}$ $GPU^{-1}$): input scaling with 512-token output (left) and sustained generation from a 1k-token prompt (right), for CENO 300M/600M/1B versus Evo 2 1B/7B and Evo 1 7B; crosses (×) mark runs that did not complete (out-of-memory). Image created with Biorender.com, with permissions.

Stage II shifted this grouped split to 28%/72%. Stage III/IV then used separate long-context weighting with a 27.75%/72.25% grouped split. These long-context stages added complete eukaryotic-genome sources sampled as long training windows (Figure 1d).

Within the shared Stage I 8k regime, loss curves followed model scale. Training and validation losses decreased across consumed tokens for all three displayed model sizes. The 1B model reached the lowest endpoint validation loss among the 300M, 600M and 1B runs (Figure 1e). These curves isolate model-size scaling within a fixed context regime.

CENO retained practical throughput at long context. In the long-prefill benchmark, output length was fixed at 512 tokens while input length increased. At 131,072 input tokens, the CENO 300M, 600M and 1B models reached 747.41, 515.83 and 315.10 generated tokens per second per GPU, respectively. Two Evo 2 baselines did not complete this input length in the same setup because of out-of-memory errors (Figure 1f; Methods 3.3). Failed comparator runs at requested lengths are retained by failure class in the source data. In sustained long-context decoding, a 1,024-token prompt was extended to 131,072 total tokens. At the terminal point, CENO 300M, 600M and 1B reached 938.97, 852.03 and 627.57 generated tokens per second per GPU, respectively. Evo 2 1B and Evo 2 7B reached 66.87 and 67.38 generated tokens per second per GPU. Together, these results establish CENO as a practical long-context backbone, supporting stronger genomic analysis, downstream applications and sequence generation.

## 2.2. Long-context CENO links distant sequence to chromatin-scale readouts

Genomic language models can now be trained on long sequence windows, but whether they use distant context in biologically relevant tasks remains a central question. Having established the CENO backbone, we tested long-context use with readouts that move from controlled sequence retrieval to chromatin-scale structure. A synthetic needle assay asked whether distant bases affected prediction. Topologically associating domain (TAD)-scale attention and annotation-linked examples tested whether long-context signals aligned with chromatin organization and regulatory elements. Frozen-state probes then asked whether boundary information was encoded in hidden representations.

We first asked whether CENO could use sequence placed far from the prediction site. In the DNA needle-in-a-haystack assay, a 100 bp sequence was inserted into a random DNA background at 10%, 50% or 90% depth and repeated at the terminal query position. Mutating the inserted needle changed

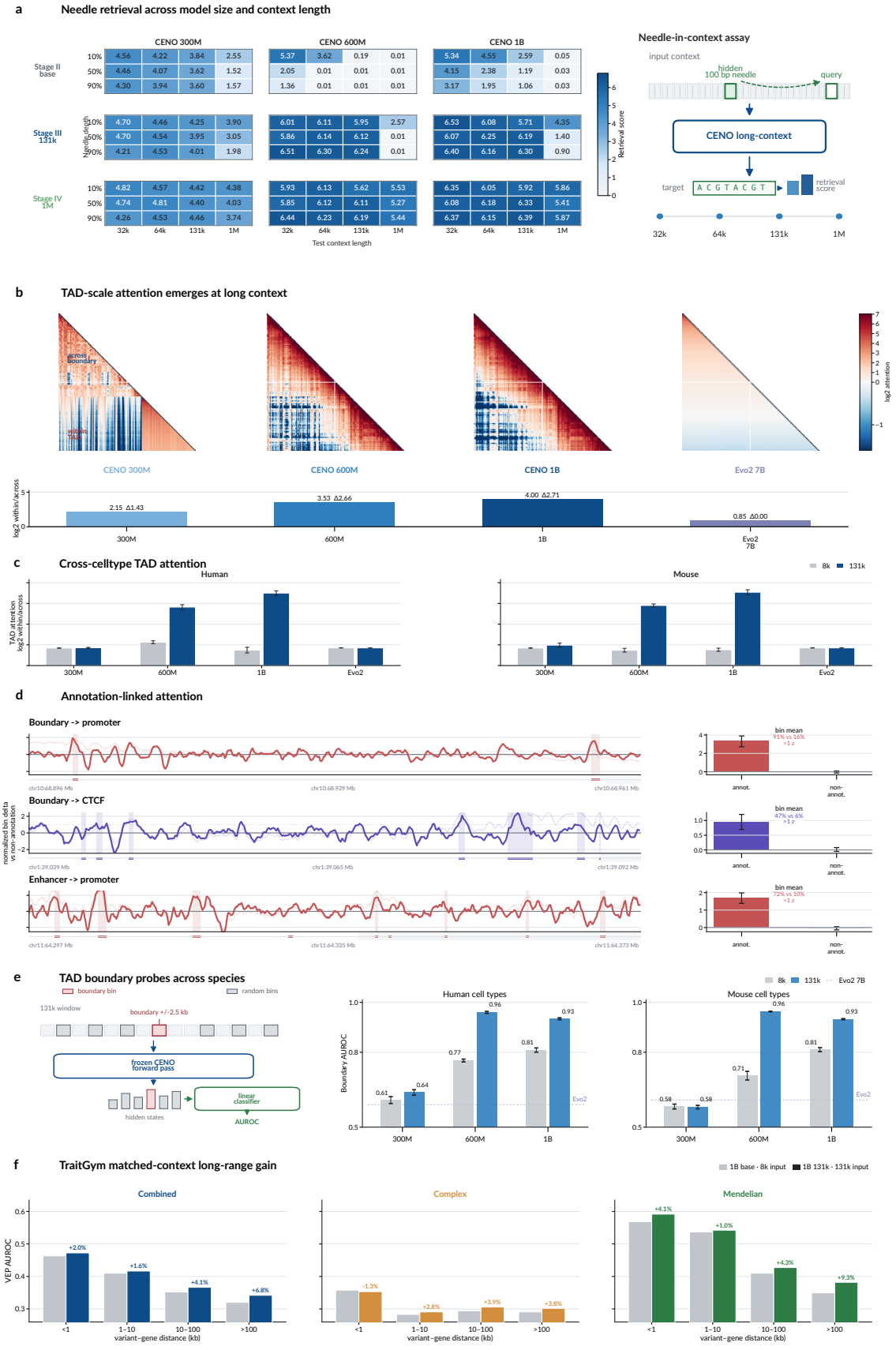

a
Needle retrieval across model size and context length
CENO 300M
CENO 600M
CENO 1B
Stage II base
Stage III 131k
Stage IV 1M
Needle depth
10%
50%
90%
4.56 4.22 3.84 2.55
4.46 4.07 3.62 1.52
4.30 3.94 3.60 1.57
5.37 3.62 0.19 0.01
2.05 0.01 0.01 0.01
1.36 0.01 0.01 0.01
5.34 4.55 2.59 0.05
4.15 2.38 1.19 0.03
3.17 1.95 1.06 0.03
4.70 4.46 4.25 3.90
4.70 4.54 3.95 3.05
4.21 4.53 4.01 1.98
6.01 6.11 5.95 2.57
5.86 6.14 6.12 0.01
6.51 6.30 6.24 0.01
6.53 6.08 5.71 4.35
6.07 6.25 6.19 1.40
6.40 6.16 6.30 0.90
4.82 4.57 4.42 4.38
4.74 4.81 4.40 4.03
4.26 4.53 4.46 3.74
5.93 6.13 5.62 5.53
5.85 6.12 6.11 5.27
6.44 6.23 6.19 5.44
6.35 6.05 5.92 5.86
6.08 6.18 6.33 5.41
6.37 6.15 6.39 5.87
32k 64k 131k 1M
Test context length
Retrieval score
Needle-in-context assay
input context
hidden
100 bp needle
query
CENO long-context
target
A C G T A C G T
retrieval score
32k 64k 131k 1M
b
TAD-scale attention emerges at long context
within
across
boundary
log2 attention
CENO 300M
CENO 600M
CENO 1B
Evo2 7B
log2 within/across
2.15 Δ1.43
3.53 Δ2.66
4.00 Δ2.71
0.85 Δ0.00
300M
600M
1B
Evo2 7B
c
Cross-celltype TAD attention
Human
Mouse
8k
131k
TAD attention log2 within/cross
300M
600M
1B
Evo2
d
Annotation-linked attention
Boundary -> promoter
Boundary -> CTCF
Enhancer -> promoter
normalized bin delta vs non-annotation
bin mean
91% vs 16%
47% vs 6%
72% vs 10%
>1 z
annot.
non-annot.
e
TAD boundary probes across species
boundary bin
random bins
131k window
boundary +/-2.5 kb
frozen CENO forward pass
linear classifier
hidden states
AUROC
Human cell types
Mouse cell types
8k
131k
Evo2 7B
Boundary AUROC
Evo2
f
TraitGym matched-context long-range gain
1B base - 8k input
1B 131k - 131k input
Combined
Complex
Mendelian
VEP AUROC
+2.0%
+1.6%
+4.1%
+6.8%
-1.3%
+2.8%
+3.9%
+3.8%
+4.1%
+1.0%
+4.3%
+9.3%
<1
1–10
10–100
>100
variant–gene distance (kb)

Figure 2 | Long-context retrieval and chromatin-scale readouts. **a,** DNA needle retrieval across 32k, 64k and 131k contexts. **b,** Matched-locus TAD attention maps and 131k/8k within/across-boundary gain; the 24-locus audit and retained examples are in the Supplementary material. **c,** Cross-cell-type TAD-attention summaries across five human cell types and five mouse cell or tissue settings. **d,** Selection-audited annotation-linked attention examples, bounded to locus-level support. **e,** Frozen-state TAD boundary probes across human and mouse settings; selected-layer, cell-level spread and label-shuffle controls are retained in source data. **f,** TraitGym matched-context long-range gain across three trait groups (Combined, Complex, Mendelian): each sub-panel pairs the 1B base scored at 8k input (grey) against the 1B Stage III model scored at 131k input (colour) by variant–gene distance, with the relative gain shown above each bin; the benefit grows at longer range and is largest for Mendelian variants beyond 100 kb.

the A/T/C/G log-probability vector at the corresponding query base. The retrieval score summarized this perturbation across the 100 needle positions (Methods 3.4).

The CENO 300M, 600M and 1B checkpoints retained retrieval scores above threshold across the 32k, 64k, 131k and 1M contexts shown in Figure 2a. Both parameter and context scaling laws hold for the task as larger model and later stages achieves higher retrieval score in longer context. This assay showed that distant sequence can affect terminal predictions.

We next asked whether model attention, without supervised boundary labels, aligned with chromatin-domain organization. TADs provide a direct readout because loci on the same side of a boundary are expected to interact more strongly than loci separated by that boundary. Because both CENO and Evo 2 contain attention layers, we directly compared their attention maps at matched TAD-boundary loci. Maps were scored as $\log_2$ within-boundary attention relative to across-boundary attention. Context scaling was assessed by comparing matched Stage II 8k and 131k windows at the same locus. In the representative locus shown in Figure 2b, 131k continuation increased this within/across-boundary score for CENO 300M, 600M and 1B by 1.43, 2.66 and 2.71 $\log_2$ units, respectively. By contrast,Evo 2 7B reference remained essentially unchanged in the same locus. We then tested whether the TAD-attention signal generalized beyond one displayed locus and one cell context. The cross-setting analysis covered five human cell types and five mouse cell or tissue settings, with three CENO model scales and an Evo 2 7B reference evaluated under matched 8k and 131k contexts. Across the five human cell types, mean $\log_2$ within/across-boundary attention increased from 1.13 to 2.82 for CENO 600M after 131k continuation. For CENO 1B, the same score increased from 0.74 to 3.50 (Figure 2c, Methods 3.5). Across five mouse cell or tissue settings, the corresponding values increased from 0.74 to 2.90 for CENO 600M and from 0.76 to 3.54 for CENO 1B. By contrast, CENO 300M and Evo 2 7B remained near baseline. Thus, the 131k continuation effect was not restricted to the representative locus, but generalized across cell-type and species settings in the 600M and 1B CENO models.

We next inspected whether high-attention regions overlapped known regulatory annotations in selected loci. This task is also zero-shot without any explicit annotation supervision. We show three CENO 1B examples: boundary-to-promoter, boundary-to-CTCF-bound candidate cis-regulatory element (cCRE) and enhancer-to-promoter attention. In these examples, annotated bins had higher normalized attention than non-annotated bins. The fractions above one normalized unit were 91% versus 16%, 47% versus 6% and 72% versus 10%, respectively (Figure 2d; Methods 3.6). These examples showed that CENO concentrates attention on functionally related sequence, linking chromatin-boundary readouts to regulatory annotations.

Finally, we tested whether chromatin-boundary information was present in hidden states of the model. A simple linear probe was trained to distinguish TAD boundary bins from matched random bins while the CENO backbone was held frozen. In human cell types, mean boundary AUROC increased from 0.767 at 8k to 0.960 at 131k for CENO 600M. For CENO 1B, the same metric increased from 0.809 at 8k to 0.934 at 131k (Figure 2e; Methods 3.7). In mouse cell or tissue settings, the corresponding increases were 0.706 to 0.963 for CENO 600M and 0.811 to 0.933 for CENO 1B. The 300M model

showed only small changes in the same probe readout.

Finally, we asked whether long-context pretraining translated into a downstream functional benefit on variant-effect prediction. On the TraitGym benchmark we compared the 1B base model scored at its 8k context against the 1B Stage III model scored at 131k input, so that model context and evaluation input length grew together, and stratified variants by variant–gene distance (Figure 2f; Methods 3.8). Matching longer context to longer input raised zero-shot AUROC in almost every distance bin, and the gain grew with distance: for the Combined trait set the relative improvement rose from +2.0% below 1 kb to +6.8% beyond 100 kb. The effect was strongest for Mendelian traits, reaching +9.3% beyond 100 kb, while Complex traits improved more modestly (+3.8% beyond 100 kb). Thus the additional context acquired during long-context pretraining yielded the largest functional gains precisely where the variant and its target gene were most distant.

Together, these assays show that long-context extension enables CENO to use distal sequence and encode chromatin-scale structure beyond one-dimensional sequence proximity.

## 2.3. Zero-shot CENO scoring across variant-effect tasks

Identifying the genetic variants that truly drive a specific biological function or disease phenotype within the vast genomic landscape remains a major challenge. Deep learning has shown considerable potential for addressing this problem. To enable a systematic evaluation of sequence models, we assembled, to our knowledge, the first comprehensive variant-effect benchmark spanning a broad range of biological processes and functional consequences.

At the molecular level, the PromoterAI and causal-eQTL datasets characterize the effects of genetic variants on RNA expression, whereas spliceVarDB evaluates variant-induced alterations in RNA splicing. The structural-variant (SV) dataset assesses whether large-scale genomic alterations lead to loss of function in specific genes. At the organismal and phenotypic levels, the BRCA1/2 and TraitGym datasets evaluate the disease-associated consequences of genetic variation. The BRCA1/2 dataset focuses on variants in breast cancer susceptibility genes, whereas TraitGym establishes causal relationships between genetic variants and both Mendelian and complex traits. In addition, we incorporated fitness datasets that provide an evolutionarily integrated measure of sequence functionality, a long-range enhancer–gene interaction dataset that captures regulatory relationships influencing gene expression, and a dataset in which different promoter–coding sequence combinations modulate naringenin production through their effects on metabolic pathways.

Beyond its broad coverage of biological functions, this benchmark is more comprehensive than conventional variant-effect datasets in three important respects. First, in addition to common single-nucleotide variants, the SV dataset includes a large collection of structural variants and their effects on gene function. Second, whereas many existing benchmarks primarily focus on coding regions, our collection contains extensive functional-effect data for noncoding regions, including ncRNA fitness, TraitGym, and SV datasets. Third, the benchmark captures regulatory effects across multiple genomic scales. The enhancer–gene dataset includes enhancer–promoter interactions spanning more than 100 kb, while the causal-eQTL dataset contains long-range associations between regulatory variants and their target genes.

Overall, this benchmark encompasses diverse biological processes, variant classes, genomic contexts, and interaction scales. It therefore provides a broad and comparatively balanced framework for evaluating and comparing the ability of genomic foundation models to understand the functional consequences of sequence variation.

We next asked whether the pretrained CENO checkpoints, before MSA post-training, could score sequence perturbations without task-specific fitting [2, 13, 7, 12]. For each variant, we constructed

matched reference and alternate windows and used the difference in sequence likelihood as a zero-shot effect score. We evaluated this score across ten benchmark families. These included splice disruption [16], BRCA1/BRCA2 variants [17, 18], regulatory variants, disease labels and expression perturbations. The remaining benchmarks covered enhancer-gene links, structural variation, molecular fitness and metabolic pathways [13, 10, 12] (Figure 3a).

The aggregate benchmark revealed a consistent signal across model scales and training stages. CENO-1B-1m and CENO-1B-131k ranked first and second, while the base checkpoint ranked fourth (Figure 3b). The 131k and 1m CENO-600M checkpoints also entered the top six. Thus, the aggregate result reflected a family-wide pattern rather than an isolated best checkpoint. It did not, however, imply that CENO led every benchmark.

Task-level results exposed this heterogeneity. In spliceVarDB, Evo2 7B base and Evo2 1B base occupied the first two positions. The three CENO-1B checkpoints followed, with summary AVG2 scores spanning 0.824–0.828 (Figure 3c). TraitGym produced a different leaderboard. Evo2 40B led, whereas the CENO-1B checkpoints formed the next cluster, with mean AUROC values spanning 0.595–0.611 (Figure 3d). CENO therefore transferred consistently, but other pretrained models retained an advantage on individual benchmarks.

Context length mattered most clearly for ClinVar. The 131k and 1m CENO-1B checkpoints gained approximately 0.04–0.05 AUROC when the input increased from 2k to 20k, placing immediately behind Evo2 40B (Figure 3e). The effect was not universal. In causal eQTLs, CENO-1B-1m performed best within 1 kb of the transcription start site, but Caduceus-ph and HyenaDNA-Medium-450k led the long-distance summary (Figure 3f). Structural-variant deletion scoring was more uniformly CENO-favourable, with all three CENO-1B stages occupying the leading AVG2 positions (Figure 3g). By contrast, metabolic-pathway performance remained distributed across CENO and GENERator checkpoints (Figure 3h).

The supplementary analyses separated broad transfer from subgroup-specific leadership (Supplementary Fig. S5). CENO-1B had the strongest family-level ClinVar score, followed by the displayed 7B group, Evo2 and CENO-600M (Supplementary Fig. S5a). Yet OmniReg-GPT led each displayed clinical-group F1 comparison, while the CENO-1B checkpoints showed marked group dependence (Supplementary Fig. S5b). The family-level advantage therefore reflected breadth across ClinVar tasks, not uniform superiority within every clinical category.

PromoterAI reinforced this distinction. AlphaGenome RNA-seq led the aggregate summary, followed by Evo2 40B. The three CENO-1B stages formed the next group, with mean AUROC values confined to 0.563–0.568 (Supplementary Fig. S5c). The same top-two separation persisted for UKBB proteome under-expression, where the CENO-1B stages clustered behind AlphaGenome RNA-seq and Evo2 40B (Supplementary Fig. S5d). These results support transfer across expression-perturbation tasks, while showing that model ordering remained task-specific.

BRCA and enhancer-gene prediction provided further boundaries. Evo2 models led both displayed BRCA panels, although CENO-1B-1m remained fourth for BRCA2 noncoding variants. All three CENO-1B stages also remained within the top eight for BRCA1 coding variants (Supplementary Fig. S5e,f). In enhancer-gene prediction, CENO-600M-1m ranked third behind Evo2 40B and Genos-10B-v2 (Supplementary Fig. S5g). Performance therefore depended on more than model scale or nominal context length.

TraitGym made this context dependence explicit (Supplementary Fig. S5h). CENO-1B-131k remained stable from 8k to 131k input, whereas CENO-1B-1m peaked at 32k and declined at 131k. The benefit of longer input therefore depended on both the checkpoint and benchmark.

Together, these analyses identify robust transfer, rather than universal dominance, as the main property

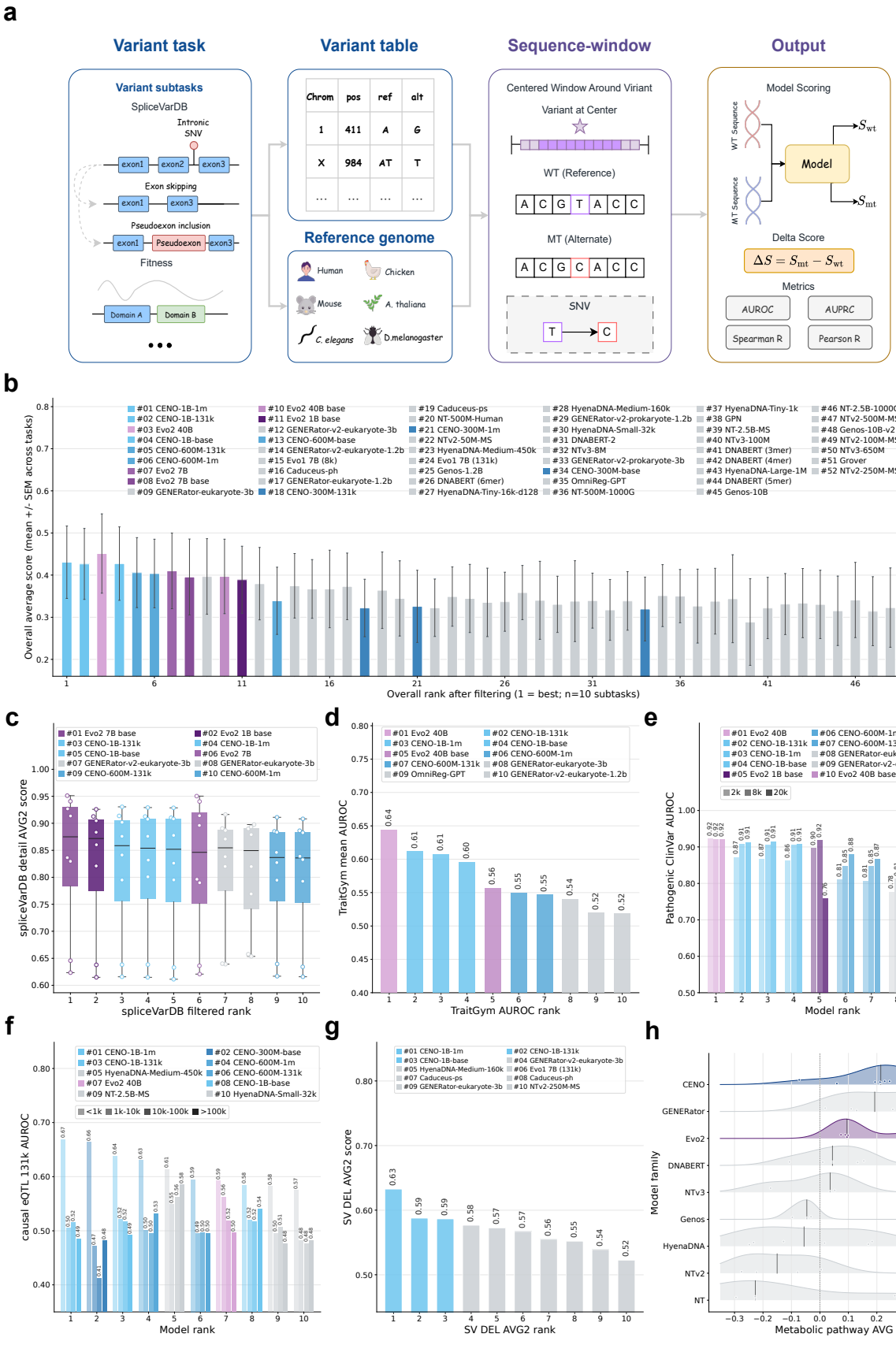

a
Variant task
Variant table
Sequence-window
Output
Variant subtasks
SpliceVarDB
Intronic SNV
exon1 exon2 exon3
Exon skipping
exon1 exon3
Pseudoexon inclusion
exon1 Pseudoexon exon3
Fitness
Domain A Domain B
Chrom pos ref alt
1 411 A G
X 984 AT T
Reference genome
Human
Chicken
Mouse
A. thaliana
C. elegans
D.melanogaster
Centered Window Around Viriant
Variant at Center
WT (Reference)
A C G T A C C
MT (Alternate)
A C G C A C C
SNV
T C
Model Scoring
WT Sequence
MT Sequence
Model
Delta Score
$\Delta S = S_{mt} - S_{wt}$
Metrics
AUROC
AUPRC
Spearman R
Pearson R
b
Overall average score (mean +/- SEM across tasks)
Overall rank after filtering (1 = best; n=10 subtasks)
c
spliceVarDB detail AVG2 score
spliceVarDB filtered rank
d
TraitGym mean AUROC
TraitGym AUROC rank
e
Pathogenic ClinVar AUROC
Model rank
f
causal eQTL 131k AUROC
Model rank
g
SV DEL AVG2 score
SV DEL AVG2 rank
h
Model family
CENO
GENERator
Evo2
DNABERT
NTv3
Genos
HyenaDNA
NTv2
NT
Metabolic pathway AVG

Figure 3 | **Cross-task zero-shot variant-effect benchmark. a,** Reference and alternate sequence scoring. Variants are represented as matched reference and alternate windows, scored by a model-derived likelihood delta and evaluated with classification or correlation metrics. **b,** Filtered overall rank across ten selected benchmark families. **c,** Distribution of spliceVarDB detail-subtask AVG2 scores for the top ten models ranked by their summary score. **d,** TraitGym mean AUROC across retained regulatory-variant subtasks. **e,** Pathogenic ClinVar AUROC by input length for the top-ranked models. **f,** Causal eQTL AUROC at 131k input, stratified by variant-to-transcription start site (TSS) distance. **g,** Structural-variant deletion AVG2 ranking. **h,** Metabolic-pathway performance summary. AVG2 denotes the mean of paired AUROC and AUPRC scores where both metrics are available.

of the pretrained CENO score. CENO remained near the leading models across heterogeneous local and long-window effects, but its relative advantage varied with assay, variant class and context. We therefore next tested two complementary capabilities: gene-scale sequence completion and the use of homologous context during MSA post-training.

## 2.4. Generation Benchmark

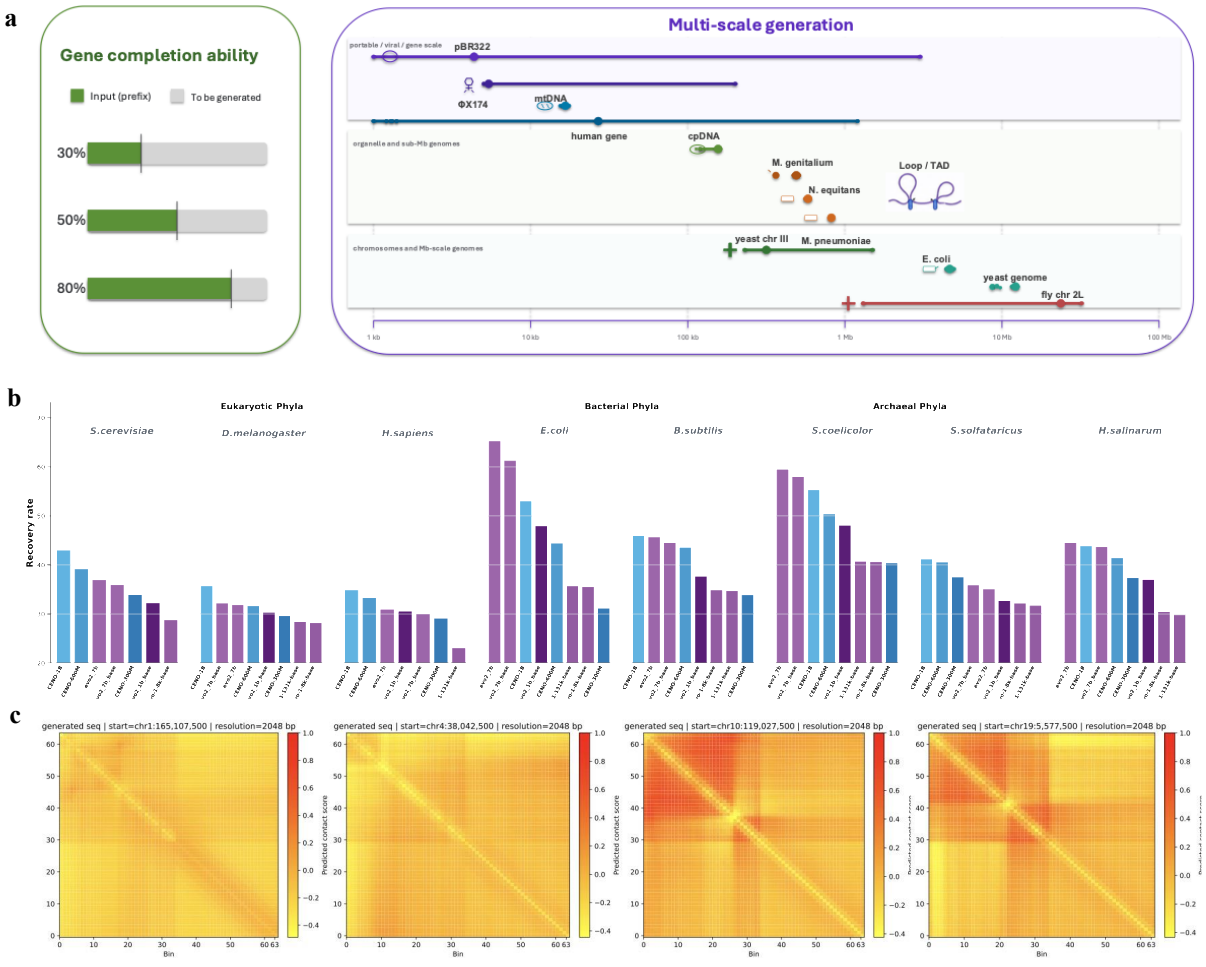
CENO is an advanced generative framework and genomic foundation model whose emergent sequence-generation capabilities have broad potential applications in genetic engineering, biopharmaceutical development, and related fields. Rather than merely reproducing naturally occurring sequences, CENO learns the underlying genomic grammar shared across diverse species and uses this knowledge to generate novel and potentially optimized DNA sequences. The biological plausibility of the generated sequences can be evaluated from multiple perspectives, including codon-usage bias, gene-coding density, gene order and synteny, and the predicted three-dimensional structures of the encoded proteins. Collectively, these analyses indicate that CENO-generated sequences can preserve key organizational and functional principles observed in natural genomes while extending beyond direct imitation of existing sequences.

CENO also demonstrates a strong capacity to recover gene sequences across diverse species. To assess this capability, 100 genes were selected from representative eukaryotic, bacterial, and archaeal species. The first 30%, 50%, or 80% of each gene sequence was provided as a prompt, and the model was tasked with completing the remaining sequence. Sequence recovery relative to the corresponding natural gene was used to quantify the model's ability to retain conserved sequence information. On this benchmark, CENO-1B achieved performance comparable to that of Evo 2-7B, despite having substantially fewer parameters. Importantly, exact recovery of the natural sequence is not the only biologically meaningful outcome, because the model may generate alternative sequences that differ from their natural counterparts while retaining equivalent or related molecular functions(Fig.4b).

More importantly, the context length of CENO can be extended from 131 kb to 1 Mb, enabling the model to capture sequence dependencies spanning most chromatin loops and a subset of topologically associating domains (TADs) in the human three-dimensional genome. This extended context provides a basis for modeling long-range regulatory interactions, including enhancer–promoter communication at well-characterized loci such as the MYC and PAX loci in GM12878 cells. CENO can consequently generate megabase-scale DNA sequences with the potential to encode higher-order chromatin organization.

To evaluate the three-dimensional organization of the generated sequences, cell-type-specific in situ Hi-C tracks predicted by AlphaGenome were used as proxies for chromatin contact profiles. Putative structural elements were further identified by scanning the generated sequences for core CTCF consensus motifs. Convergently oriented CTCF motifs were considered candidate chromatin-loop anchors, whereas clusters of CTCF-binding motifs were used to identify putative TAD boundaries. The densities of predicted loops and TADs in the generated sequences were then compared with those observed in matched natural genomic regions.

For sequence generation over a total context of 131 kb, CENO was provided with a 40-kb genomic sequence as a prompt and autoregressively generated the remaining 91 kb. The predicted contact maps of the resulting sequences contained focal interaction patterns consistent with chromatin loops(Fig.4c). These observations suggest that CENO-generated sequences can spontaneously encode sequence features compatible with higher-order chromatin folding, supporting the potential of the model to generate long genomic sequences with biologically plausible three-dimensional organization.




## 2.5. Post-Training

To adapt CENO to evolutionary context, we post-trained the pretrained long-context DNA backbones on packed examples built from real multiple sequence alignments (MSAs); we refer to the resulting MSA-adapted models as CENO-P. Each example contains a target sequence together with homologous rows. The post-training stack interleaves row-local layers, which process each sequence independently, with fusion layers that allow homologous rows to condition the target representation (Figure 5a; Methods 3.9). Variant-effect prediction then uses the same zero-shot interface as pretraining: reference and mutant alleles are scored under the same MSA context, and the likelihood delta is used as the variant score (Figure 5b; Methods 3.11). The post-training objective is a causal next-token loss over packed real MSA contexts (Methods 3.10).

We first asked whether post-training improved each backbone relative to its own pretrained checkpoint. This matched comparison isolates the effect of MSA-context adaptation from differences between model families. On BRCA variant-effect benchmarks following the GPN-MSA evaluation protocol [13],

Figure 4 | **Generation benchmark. a,**Schematic overview of sequence recovery experiments and multiscale sequence generation. **b,** Mean gene sequence recovery rates of different models across diverse species. **c,** Predicted chromatin contact map for a 131-kb sequence comprising a 40-kb genomic prompt and 91 kb of DNA autoregressively generated by CENO.

post-training increased AUPRC by 45.3%, 46.4% and 13.5% for the 300M, 600M and 1B models, respectively (Figure 5c). The corresponding BRCA2 gains were 42.6%, 43.1% and 11.3%. Thus all three CENO backbones gained variant-effect signal in their CENO-P form after exposure to real homologous context.

These gains accrue over the course of post-training rather than appearing only at the final checkpoint. Tracking BRCA1 zero-shot AUPRC over the first 20k steps with the likelihood-delta scoring interface (Methods 3.11), we found that performance climbs steeply during the first few thousand steps and then plateaus at all three scales, so most of the variant-effect signal is acquired early in post-training (Figure 5d).

We then evaluated the post-trained checkpoints on the human disease and finemapping tasks used in the GPN-Star benchmark panel [14], together with the plotted baselines (Figure 5e). These tasks comprise COSMIC somatic missense variants [19], OMIM regulatory variants using the TraitGym-processed Mendelian dataset [20, 21], and coding and noncoding GWAS fine-mapped variants from UK Biobank fine-mapping resources [22, 23, 21]. Among the models retained in this panel, the best CENO-P ranked third on OMIM (AUPRC 0.657), second on COSMIC (0.270), second on GWAS fine-mapped missense variants (0.219), and fifth on GWAS fine-mapped noncoding variants (0.173). Across these human-disease tasks CENO-P is competitive with strong specialized and conservation-based genome-wide scorers.

CENO-P's clearest advantage is on cross-species conservation. Following the GPN-Star cross-species benchmark construction [24, 25, 26, 27, 28], we scored Fisher odds-ratio enrichment on five species panels (Figure 5f; Methods 3.12). A CENO-P checkpoint ranked first among the plotted models on all five panels (1001GP, CaenDR, *D. melanogaster*, Gallus and WMGP), with odds ratios of 5.16, 8.21, 18.30, 5.16 and 2.72, respectively, surpassing genome language models an order of magnitude larger as well as alignment-based conservation scores. This uniform lead across evolutionarily distant clades shows that MSA-context post-training transfers broadly across the tree of life, and is most effective when homologous rows are available during both adaptation and variant scoring.

## 2.6. CENO enable brain cell-type-specific enhancer prediction and design

We investigated a mouse motor cortex scATAC-seq dataset for cell-type-specific enhancer prediction and design with CENO. The workflow couples a CENO-based accessibility oracle with a conditional sequence generator (Figure 6a). Candidate CREs were first derived from mouse cortex scATAC-seq peaks and organized into a cell type by accessibility matrix. The oracle was trained to predict subclass-resolved ATAC activity from the same input DNA sequence, and the resulting sequence-to-function model was used as a reward model for conditional enhancer generation. The generator was first adapted by supervised fine-tuning (SFT) to learn cell-type-conditioned CRE sequence distributions and was then further optimized by reinforcement learning (RL) to increase predicted target-cell activity while penalizing off-target activity.

We trained a CENO-based peak regression oracle by first training on 440,993 consensus peaks followed by fine-tuning on 73,323 cell-type-specific regions. To quantify oracle performance, we evaluated predictions on chromosome-held-out consensus and cell-type-specific test peaks (Figure 6b,Figure S6b). For each test region, the oracle outputs accessibility scores across mouse cortex subclasses. We flattened predictions and targets across peak-cell type pairs, excluded entries with zero observed

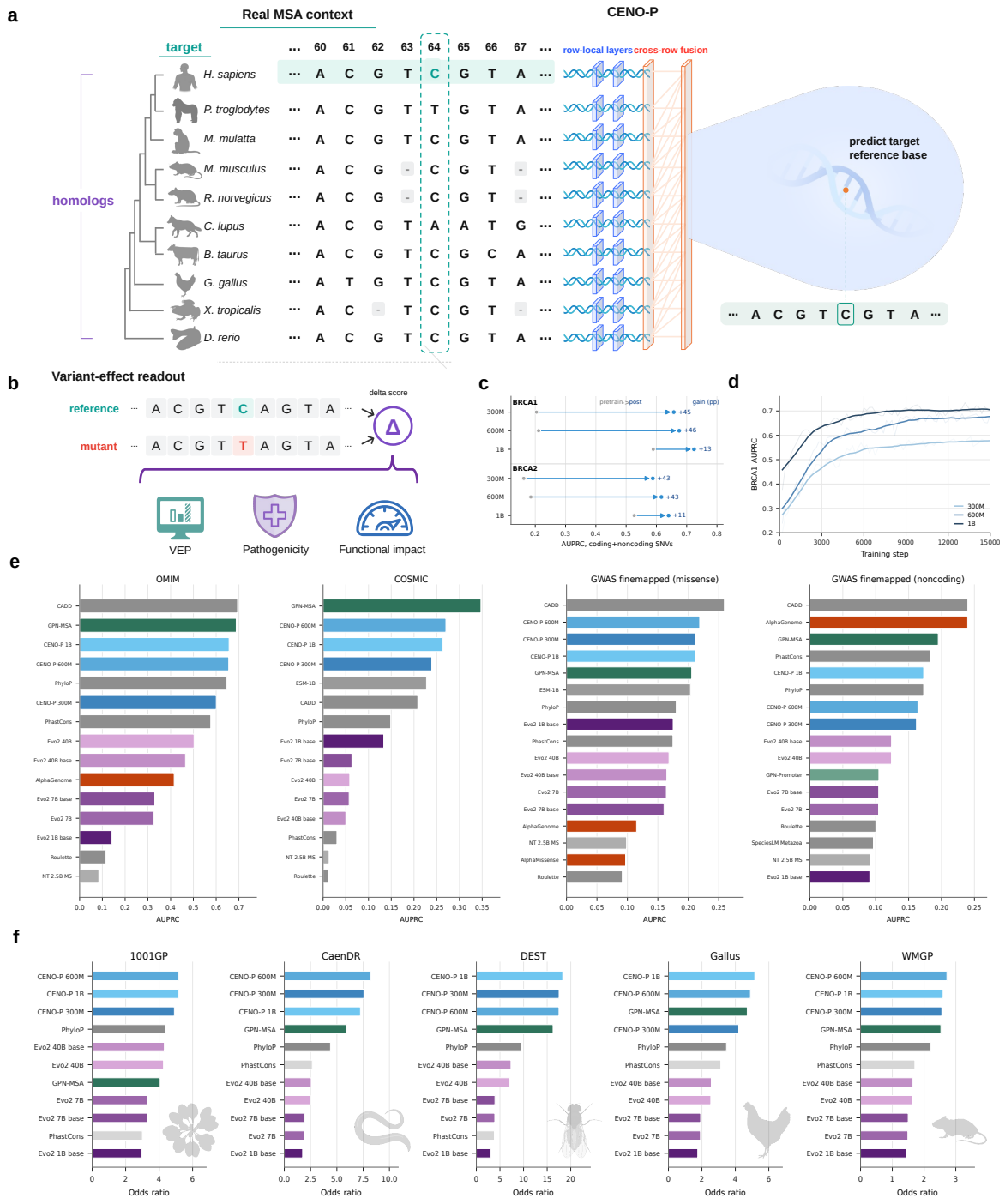
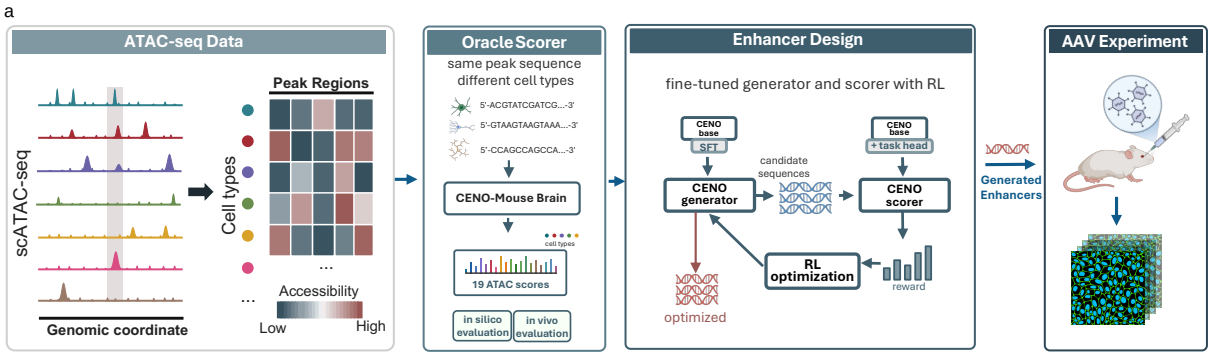
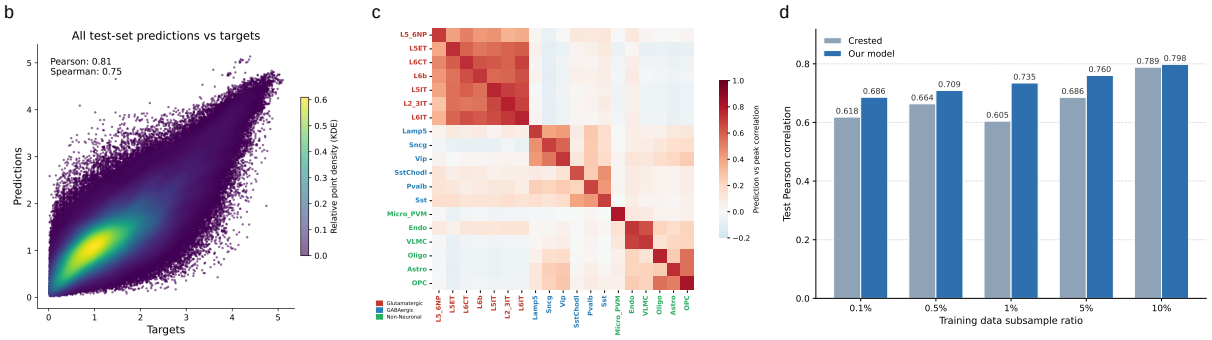
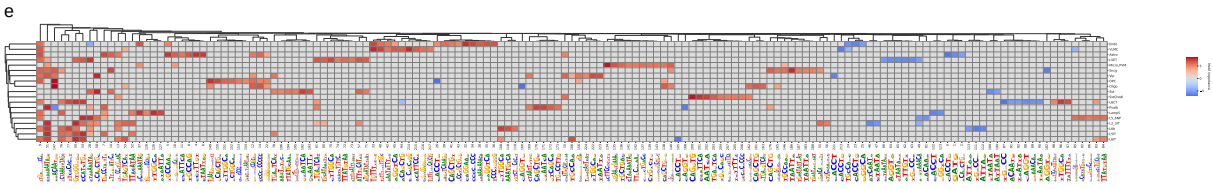
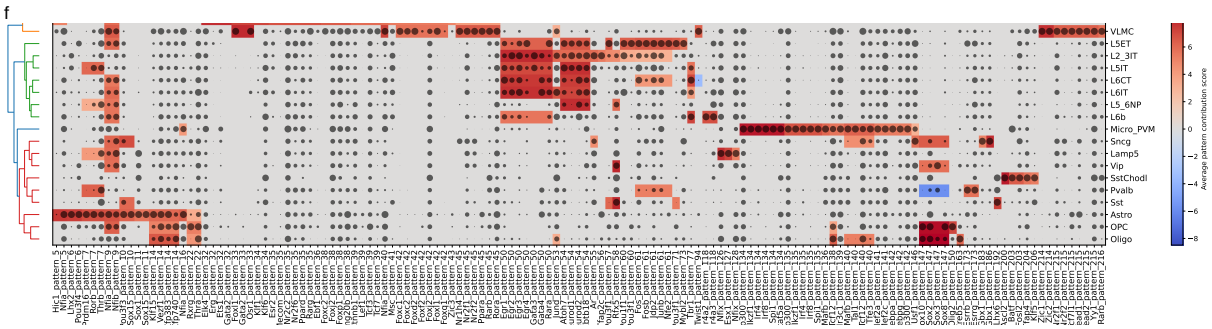



Figure 5 | CENO post-training with real MSA context. **a**, CENO packs a target sequence with homologous rows and explicit row identities. **b**, Variants are scored by the reference-versus-mutant likelihood delta under the same MSA context. **c**, Matched CENO pretraining and post-training checkpoints on BRCA1 and BRCA2. **d**, BRCA1 zero-shot AUPRC over post-training. **e**, Disease and GWAS finemapping benchmarks with all plotted models; post-trained checkpoints are labeled CENO-P. **f**, Cross-species Fisher odds-ratio enrichment benchmarks.


a
ATAC-seq Data
scATAC-seq
Peak Regions
Cell types
Accessibility
Low
High
Genomic coordinate
Oracle Scorer
same peak sequence
different cell types
5'-ACGTATCGATCG...-3'
5'-GTAAGTAAGTAAA...-3'
5'-CCAGCCAGCCA...-3'
CENO-Mouse Brain
cell types
19 ATAC scores
in silico evaluation
in vivo evaluation
Enhancer Design
fine-tuned generator and scorer with RL
CENO base
SFT
CENO generator
candidate sequences
CENO base
+ task head
CENO scorer
RL optimization
reward
optimized
Generated Enhancers
AAV Experiment


b
All test-set predictions vs targets
Pearson: 0.81
Spearman: 0.75
Predictions
Targets
Relative point density (KDE)
c
L5_6NP
L5ET
L6CT
L6b
L5IT
L2_3IT
L6IT
Lamp5
Sncg
Vip
SstChodl
Pvalb
Sst
Micro_PVM
Endo
VLMC
Oligo
Astro
OPC
Glutamatergic
GABAergic
Non-Neuronal
Prediction vs peak correlation
d
Crested
Our model
Test Pearson correlation
0.618
0.686
0.664
0.709
0.605
0.735
0.686
0.760
0.789
0.798
0.1%
0.5%
1%
5%
10%
Training data subsample ratio


e


f
VLMC
L5ET
L2_3IT
L5IT
L6CT
L6IT
L5_6NP
L6b
Micro_PVM
Sncg
Lamp5
Vip
SstChodl
Pvalb
Sst
Astro
OPC
Oligo
Average pattern contribution score

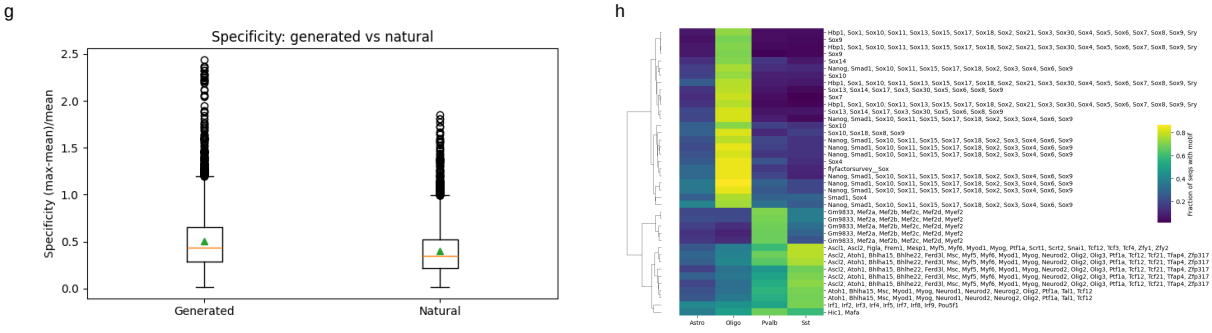

g
Specificity: generated vs natural
Specificity (max-mean)/mean
Generated
Natural
h
Astro
Oligo
Pvalb
Sst
Fraction of seqs with motif

Figure 6 | Cell-type-specific enhancer prediction and design. **a,** Overview of oracle-guided cell-type-specific CRE generation. **b,** Density scatter plot of oracle predictions and observed chromatin accessibility on held-out test peaks. **c,** Cross-cell-type correlation matrix between predicted and observed chromatin accessibility. **d,** Data-efficiency comparison between the CENO-based oracle and the CREsted baseline across different training-data subsampling ratios, evaluated by test Pearson correlation.. **e,** TF-MoDISco motif atlas derived from oracle contribution scores. **f,** Joint visualization of motif contribution scores and matched TF RNA-seq expression. **g,** Predicted specificity of generated and natural CREs. **h,** Motif enrichment across Astro, Oligo, Pvalb and Sst generated sequences.

accessibility, and compared `log1p`-transformed predictions with `log1p`-transformed peak heights. The oracle achieved a Pearson correlation of 0.81 and a Spearman correlation of 0.75 on these nonzero pairs, with the density scatter plot showing a clear monotonic relationship between predicted and observed accessibility.

We further asked whether oracle prediction channels preserve cell-type identity. For each predicted cell type and each observed cell type, we computed Pearson correlation across cell-type-specific test peaks (Figure 6c). The strongest correlations were concentrated along the diagonal, indicating that each prediction channel best matched its corresponding target cell type. Block-wise correlations within glutamatergic, GABAergic and non-neuronal groups further reflected shared accessibility programs among related cell types.

We then examined whether the oracle learned interpretable regulatory grammar. Nucleotide-level contribution scores were computed by in silico mutagenesis (ISM) , which measures the change in oracle prediction after single-nucleotide substitutions. Recurrent high-contribution sequence patterns were then identified and summarized as a cell type by motif-importance matrix for the 2,000 most cell-type-specific regions (Figure 6e). The resulting motif atlas showed structured, cell-type-associated motif usage across cortical subclasses, suggesting that oracle predictions are supported by recognizable sequence features rather than only generic accessibility-associated signals.

To connect these sequence features with cell-type biology, we jointly compared motif contribution scores with matched TF expression from RNA-seq (Figure 6f). Several high-contribution motifs showed concordant cell-type-specific TF expression patterns. For example, NFI-family motifs were prominent in astrocyte-related non-neuronal classes, whereas SOX-family motifs were enriched in Oligo/OPC-associated programs. These results indicate that the oracle captures sequence features aligned with cell-type-specific TF programs.

Having established the oracle, we next evaluated the sequences produced by oracle-guided generation. Generated sequences were sampled from the final GRPO generator, which was reinforcement-learned from the supervised fine-tuned model using the tail-only oracle MinGap reward without demonstration injection (Methods). Across all four design targets, RL-generated sequences showed higher oracle-predicted target-cell scores than natural CREs (Figure S7). We then quantified predicted cell-type specificity for generated and natural CREs (Figure 6g). Specificity was defined as the maximum predicted activity across cell types divided by the mean predicted activity across cell types, measuring whether activity is concentrated in one cell type rather than broadly distributed. Generated sequences showed an upward shift in predicted specificity relative to natural CREs, indicating that oracle-guided generation enriched for more cell-type-selective candidate enhancers.

Finally, we assessed whether generated sequences retained recognizable cell-type-associated motif grammar (Figure 6h). Motif enrichment analysis was performed on sequences generated for Astro, Oligo, Pvalb and Sst targets, and the fraction of sequences containing each motif family was summarized across target classes. Generated sequences showed target-dependent motif composition, including SOX-family enrichment in Oligo-targeted designs and MEF2-family enrichment in Pvalb-targeted designs. Thus, conditional generation increased oracle-predicted specificity while preserving

interpretable motif features associated with the target cell type.

Overall, this workflow provides a complete oracle-guided route from mouse cortex accessibility modeling to enhancer-code interpretation and cell-type-specific CRE generation, demonstrating CENO's utility for both sequence-to-function prediction and regulatory sequence design.

## 3. Discussion

CENO was developed from the view that genome modeling should become a world-modeling problem. DNA is not only a string to be classified, but a generative substrate whose local syntax, distal regulatory structure, evolutionary history and designable function are coupled across scales. In this operational sense, a genomic world model should support multiple modes of use within one backbone: reading existing sequence, scoring counterfactual variants, reconstructing or generating plausible sequence continuations, conditioning on evolutionary context and optimizing new sequences under functional constraints. CENO connects these modes through a single autoregressive backbone, long-context continuation, evolutionary MSA post-training and oracle-guided regulatory design.

The term world model should be interpreted operationally rather than metaphysically. CENO does not claim to simulate complete cellular state, organismal phenotype or physical 3D genome dynamics. Instead, it models the sequence-native world of genomic constraints: local nucleotide grammar, gene-scale organization, distal regulatory context, homologous evolutionary evidence and designable regulatory sequence space. This framing is useful because it turns a broad ambition into testable capabilities. A genomic world model should be judged by whether it can retrieve distal information, form long-context representations, score sequence interventions, recover withheld sequence, use homologous context and generate sequences under functional objectives.

The architecture and curriculum of CENO were designed to separate model capacity, context length and data composition. The hybrid Mamba–attention–MoE backbone combines efficient long sequence mixing, explicit content-based interactions and sparse expert capacity. The staged curriculum first learns broad genomic sequence grammar at 8k context, then shifts toward eukaryotic, transcript, splice and regulatory sequence, and finally extends context to 131k and 1M tokens with complete eukaryotic-genome windows. This separation is important because short-range and long-range genomic capabilities are not identical. Local motif grammar and coding syntax can often be captured from short windows, whereas distal retrieval, chromatin-boundary-associated representations and gene-scale continuation require both longer context and sufficient model capacity. In the world-model framing, increasing context length is not itself the goal; the goal is to make longer sequence context usable by the model state.

The long-context analyses indicate that CENO does more than accept longer inputs. In the DNA needle assay, mutating a distant inserted sequence changed terminal base predictions across 32k, 64k and 131k contexts, showing that distal sequence can influence the local probability distribution. In the TAD analyses, 131k continuation strengthened within-boundary relative to across-boundary attention in the 600M and 1B models and improved frozen-state boundary probes across human and mouse settings. These results suggest that long-context continuation changes the internal representation of genomic sequence, rather than merely increasing the configured context length. At the same time, these assays should be interpreted as diagnostic readouts. TAD-scale attention and linear probes provide evidence for chromatin-boundary-associated sequence representations, but they do not imply that the model reconstructs a complete physical contact map or replaces supervised 3D genome prediction.

The cross-setting behavior is an important part of this interpretation. The TAD-attention and frozen-probe signals are not restricted to one displayed locus or one cell type: they are observed across

multiple human cell types and mouse cell or tissue settings. This suggests that long-context continuation can produce reusable sequence-level boundary representations rather than a single-context artifact. However, the effect is scale dependent. The 600M and 1B models show the clearest gains, whereas the 300M model is weaker in several chromatin-boundary readouts. This pattern suggests that long-context genomic reasoning may require a capacity threshold: increasing context length alone is not sufficient unless the model has enough representational capacity to use the additional sequence.

The broader zero-shot benchmarks reinforce the need for multi-axis evaluation. Variant-effect prediction, molecular fitness, regulatory-impact tasks, structural-variant readouts and sequence generation probe different aspects of genomic modeling. CENO is competitive across many of these settings and strong in several long-context and evolutionary-context analyses, but no single benchmark family is sufficient to define general genomic ability. Human disease and fine-mapping benchmarks remain especially challenging, and task-specific or phylogeny-informed methods can remain stronger in some settings. This fragmentation should not be viewed only as a limitation of CENO. It reflects the current state of genomic foundation-model evaluation, where short-range regulatory grammar, protein or RNA fitness, noncoding variant effects, chromatin-scale structure and generation fidelity are often measured by different tasks with different inductive biases.

The world-model framing changes how genomic foundation models should be benchmarked. A model that performs well on local variant-effect prediction may still fail to use distal context. A model that accepts long inputs may still fail to generate coherent gene-scale continuations. A model that generates plausible DNA may still fail under functional design constraints. We therefore view genomic world-model evaluation as a unified benchmark paradigm rather than a single task family. Retrieval assays test whether distant sequence is accessible to the model state. Chromatin-boundary attention and frozen probes test whether long-context representations carry locus-scale structure. Variant-effect benchmarks test counterfactual perturbation sensitivity. Partial-gene continuation tests whether the learned distribution can reconstruct species-appropriate sequence organization. MSA-conditioned scoring tests evolutionary reasoning. Oracle-guided enhancer generation tests whether the same backbone can be adapted from interpretation to design.

Long-sequence generation provides a useful counterpart to variant-effect prediction. VEP asks whether a likelihood-based model assigns meaningful score changes to local perturbations of an observed reference sequence. Generation benchmarks ask a different question: given a partial gene or genomic prompt, can the model recover the withheld continuation with high similarity and species-appropriate sequence structure? This is a reconstruction-style test of gene-scale sequence understanding. It requires the model to maintain local composition, coding and noncoding grammar, species-specific sequence statistics and longer-range organization during autoregressive decoding. The improvement of generation recovery with larger CENO models and later whole-genome long-context continuation therefore supports the view that the later curriculum strengthens gene-level sequence modeling, not only inference throughput or variant scoring. This benchmark should nevertheless be interpreted as reference recovery rather than direct functional validation: recovering a natural continuation is evidence of sequence-distribution understanding, but it does not by itself prove that newly generated sequences have the intended biological function.

Evolutionary post-training adds a complementary source of information to single-sequence pretraining. By packing real MSA rows with explicit segment identities, CENO can condition the target sequence on homologous context while retaining a likelihood-delta interface for variant-effect scoring. The matched pretraining-versus-post-training comparisons on BRCA1 and BRCA2 show that the same backbone gains substantial variant-effect signal after exposure to real homologous context. The strongest pattern appears in cross-species odds-ratio-enrichment benchmarks, where CENO benefits from seeing homologous rows during both adaptation and scoring. This suggests that explicit

evolutionary context provides information that is not fully captured by isolated-sequence likelihoods or by conservation labels alone. At the same time, the benefit depends on the availability, quality and relevance of homologous alignments. Sparse clades, poorly aligned regions and species without dense alignment resources remain important future test cases.

CENO also connects interpretation to sequence design. In the mouse cortex CRE workflow, a CENO-based accessibility oracle provides a cell-type-specific sequence-to-function interface, supervised fine-tuning teaches the generative model to condition on raw text cell labels, and GRPO uses oracle feedback to shift generated sequences toward target-cell-type activity. The resulting sequences show large oracle-score shifts and retain recognizable target-associated regulatory motifs, including NFI-family motifs for astrocyte targets, SOX-family motifs for oligodendrocyte targets and MEF2-family motifs for Pvalb-associated designs. This supports the use of CENO as a design backbone rather than only a sequence scorer. Importantly, the design setting is different from the long-sequence generation benchmark. Generation recovery tests whether the pretrained model can reconstruct natural sequence continuations, whereas enhancer design tests whether the model can be guided toward new sequences satisfying a functional objective.

Oracle-guided regulatory design also introduces limitations that should be made explicit. Reinforcement learning can exploit weaknesses in the oracle, especially when generated sequences move away from the natural CRE distribution. Motif enrichment, target-associated TF programs and retrospective oracle validation provide useful evidence that the optimized sequences remain biologically plausible, but they are not substitutes for prospective experimental validation. Designed enhancers should therefore be treated as prioritized candidates for MPRA, AAV or other wet-lab assays, not as validated functional elements. Future closed-loop design should incorporate experimental feedback, distribution-shift checks, motif-level and syntax-level constraints, and negative controls designed specifically to detect oracle hacking.

Several broader limitations remain. First, the current long-context diagnostics focus on retrieval, chromatin-boundary-associated attention and TAD-boundary linear probes. These do not cover all forms of long-range genome regulation. Enhancer–promoter communication, locus control regions, insulation, structural variation and context-dependent transcriptional regulation will require additional assays and task-specific readouts. Second, downstream performance is not strictly monotonic with model size. Larger models generally benefit long-context and generation recovery, but smaller models can gain substantially from evolutionary post-training in some variant benchmarks. This suggests that data curriculum, context length, post-training objective and architecture may matter as much as total parameter count. Third, MSA post-training currently depends on real homologous rows; generated MSA contexts, low-resource species and alignment-free evolutionary conditioning remain open directions. Fourth, sequence generation benchmarks based on reference recovery measure distributional reconstruction rather than function. They are valuable for testing gene-scale sequence understanding, but should be combined with functional or experimental assays when making biological claims. Fifth, the current design workflow uses an oracle learned from available mouse-cortex accessibility data; extending genomic world models toward broader phenotype spaces will require richer oracles, perturbational measurements and prospective experimental feedback.

These limitations point to a broader path forward. Long-context genomic world models can serve as a common substrate for three increasingly functional uses: single-sequence genome modeling, evolutionary sequence interpretation and programmable sequence design. In this view, pretraining learns a short-to-long representation of genome sequence, MSA post-training adds evolutionary constraint, and oracle- or experiment-guided optimization turns the model into a design engine. Future CENO extensions could incorporate richer alignment generation, tissue- and perturbation-specific oracles, prospective validation of designed regulatory elements, and closed-loop variant-to-design workflows. More generally, our results suggest that genomic foundation models should be developed

as world models of genome sequence space: models that retain nucleotide-resolution grammar, route information across regulatory and chromatin-scale distances, recover gene-scale sequence organization across species, condition on evolutionary evidence and adapt these representations to functional design.

## Methods



### 3.1. Model architecture

CENO is a family of causal autoregressive genomic language models with a hybrid Mamba-2, attention and mixture-of-experts backbone [29, 30]. We trained three scales with approximately 300M, 600M and 1B total parameters, corresponding to approximately 100M, 200M and 400M active parameters per token. All models use a byte-level DNA tokenizer with a 512-entry vocabulary. Input and output embeddings are untied, no learned absolute position embeddings are used, and the maximum configured context length is 1,048,576 tokens.

The backbone contains Mamba-2, attention and MoE layers. Mamba-2 layers use state dimension 128 and head dimension 64. Attention layers provide explicit content-based interactions, and MoE layers use eight experts with router top-$k = 2$. The 300M, 600M and 1B models use 9, 20 and 38 layers, respectively. The 600M configuration replaces the first MoE position with a dense MLP block for training stability.

### 3.2. Pretraining data and schedules

CENO was pretrained on OpenGenome2, the same broad genomic corpus used by Evo 2 [7]. The tokenized corpus spans GTDB/IMG prokaryotic genomes, metagenomes, viral genomes, organelle genomes, eukaryotic genic windows, mRNA, splice/promoter windows, ncRNA, promoter windows and complete eukaryotic genomes. We retained the Evo 2-style train, validation and test split supplied with the tokenized corpus. Evo 2/OpenGenome2 safety exclusions, including the exclusion of eukaryotic-host viruses, were retained during preprocessing.

Pretraining used causal next-token prediction with a staged context-length curriculum. Stage I trained at 8,192 tokens with a broad OpenGenome2 mixture. Stage II kept the 8,192-token context and

increased the eukaryotic, transcript, splice and regulatory components. The 300M, 600M and 1B models used Stage I/II budgets of 1.0T/0.5T, 2.0T/1.3T and 3.5T/2.5T tokens, respectively. Stage III continued from the Stage II checkpoint at 131,072 tokens for 500B tokens. Stage IV continued from the Stage III checkpoint at 1,048,576 tokens for an additional 500B tokens. Model-specific sample counts and batch settings are listed in Supplementary Table S1.

Stage I and Stage II mixtures were selected with constrained RegMix [31]. For Stage I, 512 candidate proxy models spanning seven source categories were generated under a broad eukaryotic/prokaryotic balance constraint and evaluated on a selected subset of the VEP benchmarks. A LightGBM response model [32] was trained to map mixture weights to mean benchmark performance, scored 200,000 simulated mixtures, and produced the final Stage I weights by averaging the top 128 predicted mixtures. Stage II used the same selection procedure, using 507 candidate proxy models over nine source categories, 100,000 simulated mixtures and stronger eukaryotic, transcript, splice and regulatory weighting constraints. The fixed Stage I/II production weights are summarized in Supplementary Table S2.

Stage III and Stage IV used a long-context continuation mixture designed to preserve the Stage II eukaryotic-dominant balance while allocating part of the eukaryotic component to complete eukaryotic genomes. It retained short-window, transcript, splice, regulatory and microbial sources and added Animalia, Plantae, Fungi, Protista and Chromista complete-genome sources sampled into 131k and 1M windows. At the grouped level, Stage III/IV assigned 32.32% to prokaryotic, metagenomic, viral and organelle sources and 67.68% to eukaryotic, transcript, regulatory and complete-genome sources.

Optimization used Adam with $\beta_1 = 0.9$, $\beta_2 = 0.95$, weight decay 0.1, gradient clipping at 1.0, cosine learning-rate schedules and FP8 mixed precision. Attention and hidden dropout were set to zero. Repeat-masked lowercase bases were retained in the input sequence and assigned loss weight 0.1 in the training objective. Long-context stages used context parallelism to shard the sequence dimension. Training loss, validation loss and pretraining throughput summaries were computed directly from Megatron logs for the reported stages.

### 3.3. Inference throughput

Inference throughput was evaluated with random DNA prompts on a single GPU. Two regimes were measured. In input-length scaling, output length was fixed at 512 tokens and prompt length was varied up to 131,072 tokens where the model supported the requested context. In generation-length scaling, prompt length was fixed at 1,024 tokens and generated length was varied up to a total sequence length of 131,072 tokens. Throughput was reported as generated tokens per second per GPU for each condition. Out-of-memory or runtime failures at a requested input length were recorded as failed runs and retained in the throughput grid.

CENO checkpoints were evaluated using a modified vLLM backend with support for the CENO architecture. Evo-family baselines were evaluated with their supported inference backends and model-specific code paths on the same hardware. For the Figure 1e input-scaling display, raw benchmark summaries were retained for every condition.

### 3.4. Needle-in-a-haystack assay

We used a needle-in-a-haystack assay adapted from the Evo 2 study [7] to test whether sequence information placed far from the prediction site affected terminal predictions. For each sample, a random 100-bp DNA needle was inserted into a random DNA haystack at a specified depth, and an identical 100-bp query was appended to the end of the prompt. Main Figure 2 reports test contexts of 32k, 64k, 131k and 1M bp at insertion depths of 10%, 50% and 90%. We evaluated three

independently generated samples for every model–context–depth condition.

Retrieval was quantified using diagonal categorical-Jacobian-style perturbations. At each needle position, we first recorded the A/T/C/G log-probability vector at the aligned query position. We then substituted the inserted base with each of the other three nucleotides and recomputed the corresponding query-position vector. The position score was the mean Euclidean distance between the original and perturbed four-base vectors over the three substitutions, and the retrieval score was the mean of the 100 position scores.

In positive samples, the inserted needle was identical to the terminal query. For shuffled-null samples, the inserted needle was randomly permuted while the terminal query was held fixed. Null controls were evaluated at 10% and 100% insertion depths, with three independently generated samples per condition. Matched unshuffled records were also retained to validate each null construction but were not included in the reported null statistics. A positive trial was counted as successful when its retrieval score exceeded the prespecified threshold of 0.8. Each displayed heatmap value is the mean of three positive trials. For each model-stage and test-context setting, the supplementary summary reports nine positive records (three depths by three trials) and six shuffled-null records (two depths by three trials).

### 3.5. Zero-shot TAD attention analysis

We used TAD-boundary loci as a zero-shot readout of long-context routing. No boundary labels were used to train or adapt the models for this analysis. Source boundary files were obtained from the 4DN Project. Boundary intervals labelled Strong with score at least 0.5 were retained. We centered 131,072-bp windows on retained boundaries, sampled up to 128 positive windows per source and generated same-length negative windows at least 50,000 bp from any retained boundary window using seed 42. Negative rows were sampled from genomic windows that did not overlap the retained annotations. The main Figure 2b,c scores were computed from boundary-centered positive windows.

For CENO, maps were computed at bin resolution rather than from full token-token attention tensors. At each selected attention layer, Q and K states were mean-pooled inside 128 bp bins, converted to causal QK softmax attention per head, averaged over heads and then averaged over selected layers. Each centered matrix was partitioned into disjoint segment masks. TAD-map panels display $\log_2$ enrichment relative to the across-boundary baseline. The primary score was the $\log_2$ ratio of mean within-boundary attention to mean across-boundary attention. Context-scaling panels compared matched Stage II 8k baselines and 131k continuation settings for the same locus.

Each extracted locus has complete controls for CENO 300M, 600M and 1B Stage II 8k; CENO 300M, 600M and 1B 131k; Evo 2 7B base 8k; and Evo 2 7B 131k. To summarize cross-context behavior, we computed species-level means and s.e.m. for each model family and context length. The summaries used the within/across-boundary score across five human cell types and five mouse cell or tissue settings. Human cell types were GM12878, H1-hESC, HFFc6, IMR-90 and K562. Mouse settings were brain, cerebellar granule neurons at postnatal day 11, double-positive thymocytes, Patski cells and splenic regulatory T cells.

### 3.6. Annotation-linked attention

Annotation-linked attention was evaluated in selected CENO 1B 131k loci using cCRE, promoter, CTCF-bound, enhancer-like and CpG tracks. For each displayed case, one selected head or the median of a recorded selected-head set was aligned to genomic bins. Annotated bins were then compared with non-annotated background bins within the same local window. The track definitions were ENCODE cCRE GRCh38, UCSC CpG islands on hg38 and GENCODE v44 promoters within 2 kb of transcript

starts. Public case labels, selected layer/head indices and bin counts are retained in the Figure 2d source tables. The attention profile was averaged across query positions, divided by the median local background attention and converted to $\log_2$ observed/background units. The positive profile was then compared with a matched negative profile for the same query and target-track relationship. Profiles were smoothed over five bins, detrended by a 39-bin wide trend and robust-scaled against same-window background bins. One normalized unit denotes one background robust-scale unit above the background median after this matched-control and detrending procedure. Annotated bins are visible target track bins that exclude query bins; background bins are non-annotated, non-query bins in the same focus window. The selection audit required at least six annotated bins, at least one annotated-bin cluster and a centrality score of at least 0.45. Figure 2 reports a representative selection of these examples. The panel reports normalized attention profiles and bin-level summary bars.

## 3.7. TAD and CTCF linear probing

Frozen hidden states were extracted from each model and layer and used as fixed features for logistic-regression probes. The primary probe distinguished TAD boundary positions from matched negatives. A parallel CTCF probe distinguished CTCF motif positions from non-motif negatives and served as a local motif control. Probe positions were sampled from boundary-centered windows. Boundary positives were within 2,500 bp of the window center. Boundary negatives were at least 25,000 bp from the center and at least 200 bp from any CTCF motif center. CTCF positives used motif centers detected in each boundary-centered window, capped at 20 motifs per region. CTCF negatives were non-motif positions at least 200 bp from any motif center. Both samplers avoided the outer 64 bp of each window and used seed 2026.

Layer selection was fixed before cross-cell-type reporting. Human selected layers were chosen by H1-hESC boundary-probe performance. Mouse selected layers were chosen by brain boundary-probe performance. In-cell AUROC was the mean of five GroupKFold splits that left genomic-region groups out. Cross-cell-type transfer, where available, trained on all examples from one cell type and tested on another. Features were standardized within each training split. Logistic regression used $C = 1.0$, a maximum of 200 iterations and seed 2026. Human probe summaries used GM12878, H1-hESC, HFFc6, IMR-90 and K562. Mouse probe summaries used brain, cerebellar granule neurons at postnatal day 11, double-positive thymocytes, Patski cells and splenic regulatory T cells. Label-shuffle controls permuted labels 100 times at the selected boundary layer.

## 3.8. Matched-context variant-effect evaluation

To test whether long-context pretraining improved a downstream functional readout, we evaluated zero-shot variant-effect prediction on the TraitGym benchmark, which scores regulatory variants annotated as Mendelian or Complex traits. Variant effects were scored without task-specific training from the reference/alternate allele log-likelihood difference ($\Delta \log p$) under each model, and task performance was summarized as AUROC against the benchmark labels. Variants were stratified by variant–gene distance into <1 kb, 1–10 kb, 10–100 kb and >100 kb bins, and the Combined set pooled the Mendelian and Complex traits.

We defined a matched-context comparison in which model context and evaluation input length grew together: the 1B base (Stage II) checkpoint was scored with an 8k input window, and the 1B Stage III checkpoint was scored with a 131k input window. For each trait group and distance bin we reported both AUROC values and their relative change, $(\mathrm{AUROC}_{131k} - \mathrm{AUROC}_{8k})/\mathrm{AUROC}_{8k}$. This end-to-end contrast measures the benefit of long-context scaling rather than a fixed-input model difference; per-variant scores and additional 1M-context records are retained in the source tables.

### 3.9. MSA architecture adaptation

Post-training used a modified CENO implementation rather than the original single-stream Nemotron-H inference path. The original backbone treats the input as one causal sequence. For MSA post-training, we adapted this path to represent the input as a sequence of related rows, following the family-level modeling idea of PoET, which separates within-sequence modeling from between-sequence conditioning in an autoregressive model of homologous sequences [15]. In CENO, each nucleotide token was assigned a row identity, and the post-training stack used a layer schedule that switches between row-local processing and cross-row fusion.

In row-local layers, recurrent scans were segmented at row boundaries and attention visibility was restricted to tokens from the same row before the usual causal mask was applied. In fusion layers, the model operated on the packed causal context and could exchange information across the target and homologous rows. This gives CENO a row-local/fusion topology while keeping the pretrained hybrid Mamba–attention–MoE backbone, nucleotide vocabulary and language-modeling head unchanged.

We post-trained the 300M, 600M and 1B CENO backbones from staged checkpoints. Checkpoint names encode model scale, training stage and architecture index. The selected checkpoints used scale-specific interleaved schedules of row-local and fusion layers (Table S4).

### 3.10. MSA-context post-training

We adapted pretrained CENO checkpoints with an alignment-conditioned training signal derived from GPN-MSA [13]. The key transfer was the source of conditioning, not the exact prediction objective. GPN-MSA uses an alignment-conditioned masked nucleotide prediction objective, in which the masked target position can use both flanking sequence and homologous alignment columns. CENO remains a causal autoregressive model, so we converted this idea into next-token prediction on packed real MSA contexts. Each training example consisted of a target reference sequence and homologous rows sampled from the human multispecies alignment store.

Let $x = (x_1, \ldots, x_T)$ denote the packed MSA sequence after concatenating the target and homologous rows, and let $w_t$ denote the training weight for position $t$. Post-training minimized a weighted causal negative log-likelihood,

$$\mathcal{L}_{\mathrm{MSA}} = -\frac{\sum_{t=1}^{T-1} w_{t+1} \log p_\theta(x_{t+1} \mid x_{\le t})}{\sum_{t=1}^{T-1} w_{t+1}}.$$

Padding positions and row-boundary transitions were excluded from the loss, so the model was not trained to treat the last nucleotide of one row as the biological predecessor of the first nucleotide of the next. No mask token was inserted, and no bidirectional prediction loss was used. Thus the post-training objective changes the causal sequence likelihood used for downstream variant-effect prediction, but does not introduce a supervised variant classifier or task-specific decision boundary.

The target sequence and homologous rows were concatenated into a single packed sequence before being passed to the model. The row-local/fusion pattern described above interleaves within-row nucleotide modeling with cross-row context exchange in a single causal stack.

### 3.11. Zero-shot variant scoring after post-training

Post-trained CENO models were evaluated with a likelihood-delta interface matching the alternate-versus-reference scoring principle used by GPN-MSA [13]. For each variant, we constructed matched packed contexts for the reference and mutant alleles, keeping the MSA-derived homologous rows fixed across the two scores. Let $C$ denote this fixed homologous context and let $y$ denote the target

row. We scored each target row by its mean causal log-likelihood,

$$\ell_\theta(y; C) = \frac{1}{|\mathcal{T}|} \sum_{t \in \mathcal{T}} \log p_\theta(y_t \mid C, y_{<t}),$$

where $\mathcal{T}$ contains the scored target-row positions. The variant delta was then

$$\Delta(v) = \ell_\theta(y^{\text{mut}}; C) - \ell_\theta(y^{\text{ref}}; C).$$

Thus negative $\Delta(v)$ values indicate that the mutant sequence is less likely than the reference sequence under the same MSA context. Binary classification metrics use $-\Delta(v)$ as the deleteriousness score, whereas odds-ratio analyses use the signed delta directly with lower scores treated as more deleterious.

### 3.12. Cross-species Fisher odds ratios

For the cross-species enrichment benchmarks, all scores were oriented so that lower values indicate stronger predicted deleteriousness. For CENO this score is the signed likelihood delta $\Delta(v)$; for conservation baselines with the opposite native direction, scores were negated before thresholding. Let $s_i$ denote the oriented score and $y_i \in \{0, 1\}$ the task label, with $y_i = 1$ for the positive or rare class and $y_i = 0$ for the negative or common class. For each model and task, the cutoff was set from the negative class,

$$\tau_p = Q_p\left(\{s_i : y_i = 0\}\right),$$

with $p = 1\%$ in Figure 5f. The low-score tier was defined by $s_i \le \tau_p$. We then counted

$$n_{a,\le} = \#\{i : y_i = a,\ s_i \le \tau_p\}, \qquad n_{a,>} = \#\{i : y_i = a,\ s_i > \tau_p\},$$

formed the contingency table

$$\begin{pmatrix} n_{1,\le} & n_{1,>} \\ n_{0,\le} & n_{0,>} \end{pmatrix},$$

and applied a one-sided Fisher exact test with alternative “greater”. The reported odds ratio was

$$\text{OR}_p = \frac{n_{1,\le}/n_{1,>}}{n_{0,\le}/n_{0,>}}.$$

Thus $\text{OR}_p > 1$ indicates enrichment of positive or rare variants in the lowest-scoring tier after controlling the negative-class threshold.

### 3.13. Cell-type-specific Enhancer Prediction

The cell-type-specific cis-regulatory element (CRE) prediction and design workflow consists of two components. The first component is a sequence-to-accessibility oracle that predicts chromatin accessibility across mouse cortex subclasses from DNA sequence. The second component is a conditional enhancer generator that uses the oracle as a sequence-level scoring function during supervised fine-tuning and reinforcement learning. The oracle and the generator share the same pretrained DNA language-model backbone but are trained for different objectives. The same frozen oracle is used consistently to filter supervised fine-tuning examples, to define the reinforcement-learning reward, and to score generated sequences, so that data selection, optimization and evaluation share one accessibility model.

**Cell-type-specific accessibility oracle** The oracle takes a 2,114 bp DNA sequence as input and predicts accessibility scores across 19 mouse cortex subclasses. The model uses the 600M CENO DNA language-model checkpoint as an unfrozen backbone. DNA sequences are tokenized with the byte-level DNA tokenizer and passed through the backbone to obtain last-layer hidden states. For an input batch, the backbone produces hidden representations with shape $[B, 2114, 1024]$. These representations are then used as input channels for a dilated convolutional neural network head, followed by pooling and a linear output layer that produces a $[B, 19]$ vector of subclass-resolved accessibility predictions.

The oracle was trained using mouse cortex ATAC-seq dataset from the BICCN competition. Candidate CRE regions were represented by fixed 2,114 bp genomic windows centered on the corresponding peak regions. The target matrix contains accessibility values for the 19 mouse cortex subclasses. To reduce sequence leakage between training and evaluation, we used a chromosome-level split: chromosomes 8 and 10 were used for validation, chromosomes 9 and 18 were used for testing, and the remaining chromosomes were used for training.

Oracle training was performed in two stages. In the first stage, the model was trained on all consensus peaks to learn a general sequence-to-accessibility mapping. In the second stage, the model was further trained on cell-type-specific peak regions, which are preferentially accessible in particular mouse cortex subclasses. This two-stage strategy was used to retain broad accessibility prediction while increasing emphasis on cell-type-specific signals.

The training objective combines a pattern-matching term and a magnitude-matching term:

$$\mathcal{L} = -\cos(\hat{\mathbf{y}}, \mathbf{y}) + \mathrm{MSE}\left(\log(1+\hat{\mathbf{y}}), \log(1+\mathbf{y})\right),$$

where $\hat{\mathbf{y}}$ is the predicted accessibility vector and $\mathbf{y}$ is the target accessibility vector. The cosine term compares the relative accessibility pattern across cell types, whereas the log-MSE term penalizes differences in signal magnitude after log transformation. Training used Adam optimization, learning-rate reduction on validation plateau, early stopping, reverse-complement augmentation and small stochastic sequence shifts.

**In silico oracle evaluation** The oracle was evaluated on chromosome-held-out test regions, including both consensus peaks and cell-type-specific peaks. Prediction accuracy was quantified by comparing predicted and observed accessibility values across test regions and cell types using Pearson correlation, Spearman correlation and related regression metrics.

**In vivo oracle evaluation** We further evaluated the oracle using a retrospective set of AAV-tested cortical enhancers. The enhancer set contains mouse-derived candidates in mm10 coordinates and human-derived candidates with provided mm10 orthologous liftover coordinates. Mouse-derived enhancers were used with their original mm10 coordinates. Human-derived enhancers were mapped to mm10 using the provided orthologous coordinates, and entries without an orthologous mapping were excluded.

Because the in vivo enhancer annotations and the oracle output use different cell-type taxonomies, the reported in vivo categories were mapped to the 19 oracle subclasses before evaluation. For each enhancer, oracle scores were computed across the mapped subclasses. Candidate enhancers were ranked by the oracle score for the corresponding target class.

Retrospective in vivo evaluation used complementary readouts. Epifluorescence labels were converted into a rank-based metric using the reported enhancer category and signal strength. SSv4 single-cell quantification was used as an additional metric by weighting enhancer categories with the measured

cell-type fraction. We also computed precision, recall and F1 in a multi-label setting, using the reported active cell type or cell types as ground truth. In addition, normalized enrichment score (NES) was calculated for enhancer categories in the ranked list to quantify category enrichment among high-ranked candidates.

**Conditional supervised fine-tuning for CRE generation** The conditional enhancer generator was initialized from the 600M CENO DNA language-model checkpoint. Supervised fine-tuning was used to adapt the model from general DNA sequence modeling to conditional CRE generation. Cell-type conditions were represented as single raw-text prefix tokens. Because the tokenizer operates at the byte level, cell-type labels for the four design targets, `Astro`, `Oligo`, `Pvalb` and `Sst`, were mapped to byte slots that are unused by the nucleotide alphabet, so they were encoded directly without modifying the vocabulary or the embedding table.

Each supervised fine-tuning example consisted of a single cell-type prefix token followed by a target CRE sequence, denoted schematically as

$$\texttt{<Astro>}\,\langle \text{500 bp sequence} \rangle.$$

The supervised objective was autoregressive next-token prediction over the CRE sequence conditioned on the prefix. Under this formulation, the model learns the conditional sequence distribution

$$p(\mathbf{x}_{1:L} \mid \text{cell-type label}),$$

where $\mathbf{x}_{1:L}$ is a 500 bp CRE sequence.

The supervised fine-tuning dataset was constructed from mouse cortex candidate CREs. For each candidate region, a 2,114 bp window was extracted from the mm10 reference genome, and the geometric-center 500 bp of this window was used as the generation target; regions whose center 500 bp contained characters outside A/C/G/T were removed. For each of the four design targets, candidate CREs were scored with the frozen oracle, and a target-specific MinGap score was computed over the four design classes:

$$\text{MinGap}(c) = s_c - \max_{c' \neq c} s_{c'},$$

where $s_c$ is the oracle score for the target cell type and the maximum is taken over the other three design classes. CREs with MinGap $\geq 5$ were retained as high-specificity training examples, and up to a fixed per-cell number of top-scoring regions was kept for each target class (8,000 for training and 1,000 for validation), with training and validation subsets selected separately per class. Reverse-complement augmentation was applied during data preparation.

To emphasize the most cell-type-selective examples without discarding distributional diversity, the per-token cross-entropy loss was weighted by a monotone function of the example MinGap,

$$w(c) = \text{clip}\left(\frac{\text{MinGap}(c) - m_0}{s_0} + 1,\ w_{\min},\ w_{\max}\right),$$

with $m_0 = 10$, $s_0 = 10$, $w_{\min} = 0.5$ and $w_{\max} = 2.0$, so that an example at MinGap = 10 receives unit weight, high-specificity examples are up-weighted up to 2×, and the least specific retained examples are mildly down-weighted. Supervised fine-tuning used the AdamW optimizer with a cosine schedule for six epochs.

**Oracle-guided reinforcement learning** After supervised fine-tuning, the generator was further optimized with oracle-guided reinforcement learning, initialized from the supervised fine-tuned

checkpoint. Optimization used group relative policy optimization (GRPO) with the decoupled-clip and dynamic-sampling (DAPO) objective. During this stage, each prompt was the single cell-type prefix token, and the model autoregressively generated a 500 bp DNA sequence. Generated sequences were filtered to retain valid DNA sequences, center-padded with `N` to the 2,114 bp oracle input length, and scored by the frozen oracle.

The base reward for a generated sequence $\mathbf{x}$ and target cell type $c$ was the oracle MinGap over the four design classes,

$$\mathrm{MinGap}(\mathbf{x}, c) = s_c(\mathbf{x}) - \max_{c' \neq c} s_{c'}(\mathbf{x}),$$

where $s_c(\mathbf{x})$ is the oracle-predicted activity for the target cell type and the maximum is taken over the other three design classes. To concentrate the learning signal on high-specificity completions, we used a tail-only shaping of this quantity,

$$r(\mathbf{x}, c) = \max\big(0,\ \mathrm{MinGap}(\mathbf{x}, c)\big),$$

and assigned a fixed negative reward to sequences that failed the validity filter. During GRPO training, multiple completions were sampled from the same cell-type prompt and optimized using relative reward differences within the sampled group. Training used 32 completions per prompt, sampling temperature 0.9, a KL-regularization coefficient of 0.03 toward the supervised fine-tuned reference, group-level reward scaling, and was run for 300 optimization steps; no demonstration sequences were injected into the sampled groups. The final generator was selected from this run. Supervised fine-tuning and reinforcement-learning hyperparameters are summarized in Supplementary Table S5.

## References

[1] Zhihan Zhou et al. “DNABERT-2: Efficient Foundation Model and Benchmark For Multi-Species Genomes”. In: *The Twelfth International Conference on Learning Representations*. 2024. URL: https://proceedings.iclr.cc/paper_files/paper/2024/hash/b633e7052970b8f5aa1a69164d99e9e8-Abstract-Conference.html.

[2] Hugo Dalla-Torre et al. “Nucleotide Transformer: building and evaluating robust foundation models for human genomics”. In: *Nature Methods* 22.2 (2025), pp. 287–297. DOI: 10.1038/s41592-024-02523-z. URL: https://www.nature.com/articles/s41592-024-02523-z.

[3] Peng Ye et al. “Genomics-fm: Universal foundation model for versatile and data-efficient functional genomic analysis”. In: *bioRxiv* (2024), pp. 2024–07.

[4] Eric Nguyen et al. “HyenaDNA: Long-range genomic sequence modeling at single nucleotide resolution”. In: *Advances in Neural Information Processing Systems*. Vol. 36. Curran Associates, Inc., 2023, pp. 43177–43201. URL: https://proceedings.neurips.cc/paper_files/paper/2023/hash/86ab6927ee4ae9bde4247793c46797c7-Abstract-Conference.html.

[5] Yair Schiff et al. “Caduceus: Bi-Directional Equivariant Long-Range DNA Sequence Modeling”. In: *Proceedings of the 41st International Conference on Machine Learning*. Vol. 235. Proceedings of Machine Learning Research. PMLR, 2024, pp. 43632–43648. URL: https://proceedings.mlr.press/v235/schiff24a.html.

[6] Eric Nguyen et al. “Sequence modeling and design from molecular to genome scale with Evo”. In: *Science* 386.6723 (2024), eado9336. DOI: 10.1126/science.ado9336. URL: https://www.science.org/doi/10.1126/science.ado9336.

[7] Garyk Brixi et al. “Genome modelling and design across all domains of life with Evo 2”. In: *Nature* 652.8112 (2026), pp. 1349–1361. DOI: 10.1038/s41586-026-10176-5. URL: https://www.nature.com/articles/s41586-026-10176-5.

[8] Mingqian Ma et al. “Hybridna: A hybrid transformer-mamba2 long-range dna language model”. In: *arXiv preprint arXiv:2502.10807* (2025).

[9] Loubna Ben Allal et al. “Carbon: Decoding the Language of Life”. In: *bioRxiv* (2026), pp. 2026–05.

[10] Ziga Avsec et al. “Effective gene expression prediction from sequence by integrating long-range interactions”. In: *Nature Methods* 18 (2021), pp. 1196–1203. DOI: 10.1038/s41592-021-01252-x. URL: https://www.nature.com/articles/s41592-021-01252-x.

[11] Johannes Linder et al. “Predicting RNA-seq coverage from DNA sequence as a unifying model of gene regulation”. In: *Nature Genetics* 57.4 (2025), pp. 949–961. DOI: 10.1038/s41588-024-02053-6. URL: https://www.nature.com/articles/s41588-024-02053-6.

[12] Žiga Avsec et al. “Advancing regulatory variant effect prediction with AlphaGenome”. In: *Nature* 649.8099 (2026), pp. 1206–1218. DOI: 10.1038/s41586-025-10014-0. URL: https://www.nature.com/articles/s41586-025-10014-0.

[13] Gonzalo Benegas et al. “A DNA language model based on multispecies alignment predicts the effects of genome-wide variants”. In: *Nature Biotechnology* 43.12 (2025), pp. 1960–1965. DOI: 10.1038/s41587-024-02511-w.

[14] Chengzhong Ye et al. “Predicting functional constraints across evolutionary timescales with phylogeny-informed genomic language models”. In: *bioRxiv* (2025).

[15] Timothy Truong Jr and Tristan Bepler. “Poet: A generative model of protein families as sequences-of-sequences”. In: *Advances in Neural Information Processing Systems* 36 (2023), pp. 77379–77415.

[16] Kishore Jaganathan et al. “Predicting splicing from primary sequence with deep learning”. In: *Cell* 176.3 (2019), pp. 535–548.

[17] Gregory M Findlay et al. “Accurate classification of BRCA1 variants with saturation genome editing”. In: *Nature* 562.7726 (2018), pp. 217–222.

[18] Sounak Sahu et al. “Saturation genome editing-based clinical classification of BRCA2 variants”. In: *Nature* 638.8050 (2025), pp. 538–545.

[19] John G Tate et al. “COSMIC: the Catalogue Of Somatic Mutations In Cancer”. In: *Nucleic Acids Research* 47.D1 (2019), pp. D941–D947. ISSN: 0305-1048. DOI: 10.1093/nar/gky1015. eprint: https://academic.oup.com/nar/article-pdf/47/D1/D941/27441712/gky1015.pdf. URL: https://doi.org/10.1093/nar/gky1015.

[20] Joanna S. Amberger et al. “OMIM.org: Online Mendelian Inheritance in Man (OMIM®), an online catalog of human genes and genetic disorders”. In: *Nucleic Acids Research* 43.D1 (2015), pp. D789–D798. ISSN: 0305-1048. DOI: 10.1093/nar/gku1205. eprint: https://academic.oup.com/nar/article-pdf/43/D1/D789/7329486/gku1205.pdf. URL: https://doi.org/10.1093/nar/gku1205.

[21] Gonzalo Benegas, Gökcen Eraslan, and Yun S. Song. “Benchmarking DNA Sequence Models for Causal Regulatory Variant Prediction in Human Genetics”. In: *bioRxiv* (2025). DOI: 10.1101/2025.02.11.637758. eprint: https://www.biorxiv.org/content/early/2025/02/12/2025.02.11.637758.full.pdf. URL: https://www.biorxiv.org/content/early/2025/02/12/2025.02.11.637758.

[22] Masahiro Kanai et al. “Insights from complex trait fine-mapping across diverse populations”. In: *medRxiv* (2021). DOI: 10.1101/2021.09.03.21262975. eprint: https://www.medrxiv.org/content/early/2021/09/05/2021.09.03.21262975.full.pdf. URL: https://www.medrxiv.org/content/early/2021/09/05/2021.09.03.21262975.

[23] Clare Bycroft et al. “The UK Biobank resource with deep phenotyping and genomic data”. In: *Nature* 562 (2018), pp. 203–209. URL: https://doi.org/10.1038/s41586-018-0579-z.

[24] Carlos Alonso-Blanco et al. “1,135 Genomes Reveal the Global Pattern of Polymorphism in Arabidopsis thaliana”. In: *Cell* 166.2 (2016), pp. 481–491. ISSN: 0092-8674. DOI: https://doi.org/10.1016/j.cell.2016.05.063. URL: https://www.sciencedirect.com/science/article/pii/S0092867416306675.

[25] Timothy A Crombie et al. “CaeNDR, the Caenorhabditis Natural Diversity Resource”. In: *Nucleic Acids Research* 52.D1 (2024), pp. D850–D858. ISSN: 0305-1048. DOI: 10.1093/nar/gkad887. eprint: https://academic.oup.com/nar/article-pdf/52/D1/D850/55039590/gkad887.pdf. URL: https://doi.org/10.1093/nar/gkad887.

[26] Joaquin C B Nunez et al. “Footprints of Worldwide Adaptation in Structured Populations of Drosophila melanogaster Through the Expanded DEST 2.0 Genomic Resource”. In: *Molecular Biology and Evolution* 42.8 (2025), msaf132. ISSN: 1537-1719. DOI: 10.1093/molbev/msaf132. eprint: https://academic.oup.com/mbe/article-pdf/42/8/msaf132/64079027/msaf132.pdf. URL: https://doi.org/10.1093/molbev/msaf132.

[27] Weiwei Fu et al. “Galbase: A Comprehensive Repository for Integrating Chicken Multi-Omics Data”. In: *BMC Genomics* 23.1 (2022), p. 364. DOI: 10.1186/s12864-022-08598-2. URL: https://doi.org/10.1186/s12864-022-08598-2.

[28] Bettina Harr et al. “Genomic resources for wild populations of the house mouse, Mus musculus and its close relative Mus spretus”. In: *Scientific Data* 3.1 (2016), p. 160075. DOI: 10.1038/sdata.2016.75. URL: https://doi.org/10.1038/sdata.2016.75.

[29] Tri Dao and Albert Gu. “Transformers are SSMs: Generalized Models and Efficient Algorithms Through Structured State Space Duality”. In: *Proceedings of the 41st International Conference on Machine Learning*. Vol. 235. Proceedings of Machine Learning Research. PMLR, 2024, pp. 10041–10071. URL: https://proceedings.mlr.press/v235/dao24a.html.

[30] William Fedus, Barret Zoph, and Noam Shazeer. “Switch Transformers: Scaling to trillion parameter models with simple and efficient sparsity”. In: *Journal of Machine Learning Research* 23.120 (2022), pp. 1–39. URL: https://jmlr.org/papers/v23/21-0998.html.

[31] Qian Liu et al. “RegMix: Data Mixture as Regression for Language Model Pre-training”. In: *The Thirteenth International Conference on Learning Representations*. 2025. URL: https://proceedings.iclr.cc/paper_files/paper/2025/hash/5f67d864aae6115374fed7beddd119e0-Abstract-Conference.html.

[32] Guolin Ke et al. “LightGBM: A Highly Efficient Gradient Boosting Decision Tree”. In: *Advances in Neural Information Processing Systems*. Vol. 30. Curran Associates, Inc., 2017, pp. 3149–3157. URL: https://papers.nips.cc/paper/6907-lightgbm-a-highly-efficient-gradient-boosting-decision-tree.

# Contributions

### Core Authors

Mingqian Ma[1,*], Yucheng Wu[1,2,*], Xin Chen[1,3,*], Feifei Jiang[1,4,*], Peijun Lin[5], Dongxin Ye[6,7]

### Scientific Directors

Yidi Sun[8], Yijing Zhang[2], Tianqiong Shi[9], Yu Zhao[10], Wanli Ouyang[11,12,1], Bowen Zhou[1,13]

### Corresponding Authors

Lei Bai[1], Yuchen Ren[1]

# Affiliation

[1]Shanghai Artificial Intelligence Laboratory

[2]Fudan University

[3]The University of Sydney

[4]Shanghai Jiao Tong University

[5]University of Southern California

[6]Shanghai Innovation Institute

[7]University of Electronic Science and Technology of China

[8]Center for Excellence in Brain Science and Intelligence Technology, Chinese Academy of Sciences

[9]Nanjing Normal University

[10]Westlake University

[11]Shenzhen Loop Area Institute

[12]The Chinese University of Hong Kong

[13]Tsinghua University

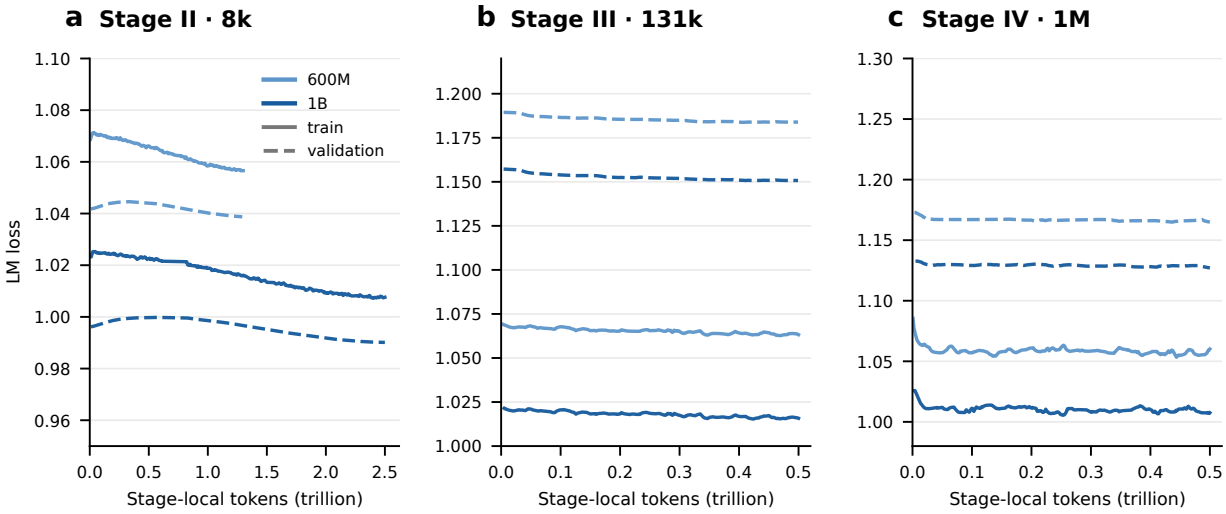
# CENO

*Supplementary Material*

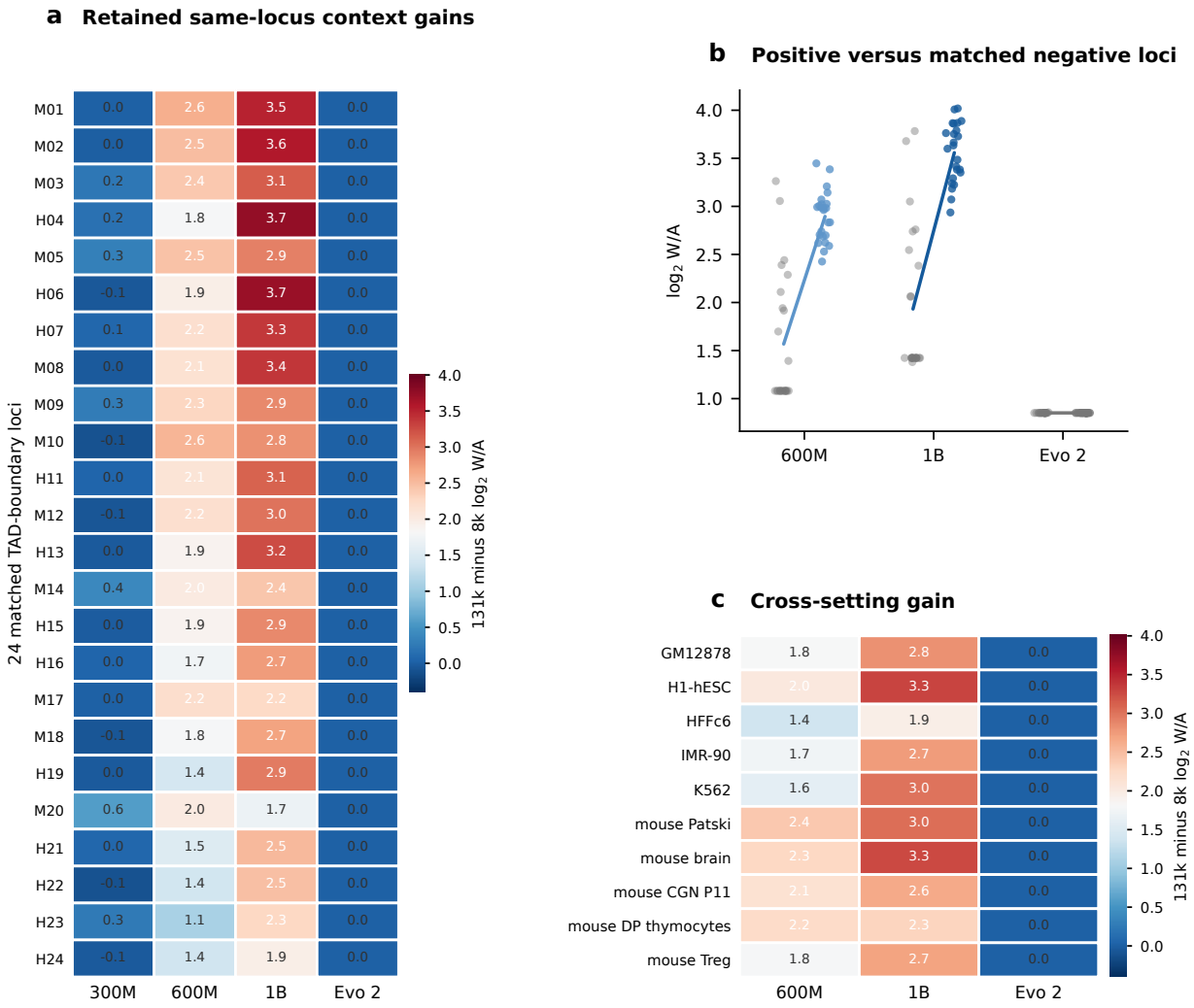
---

## Supplementary contents

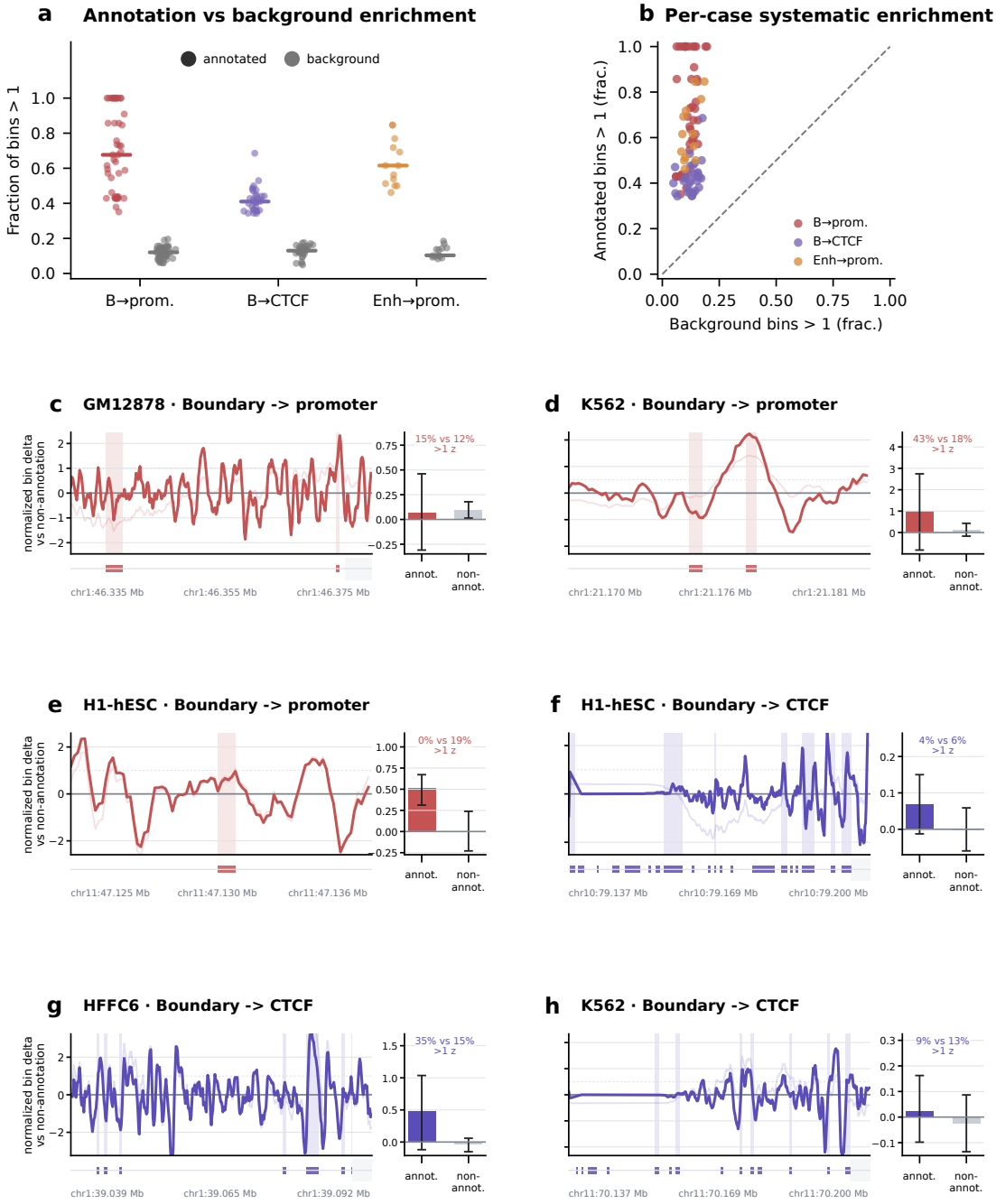
# A. Supplementary Figures



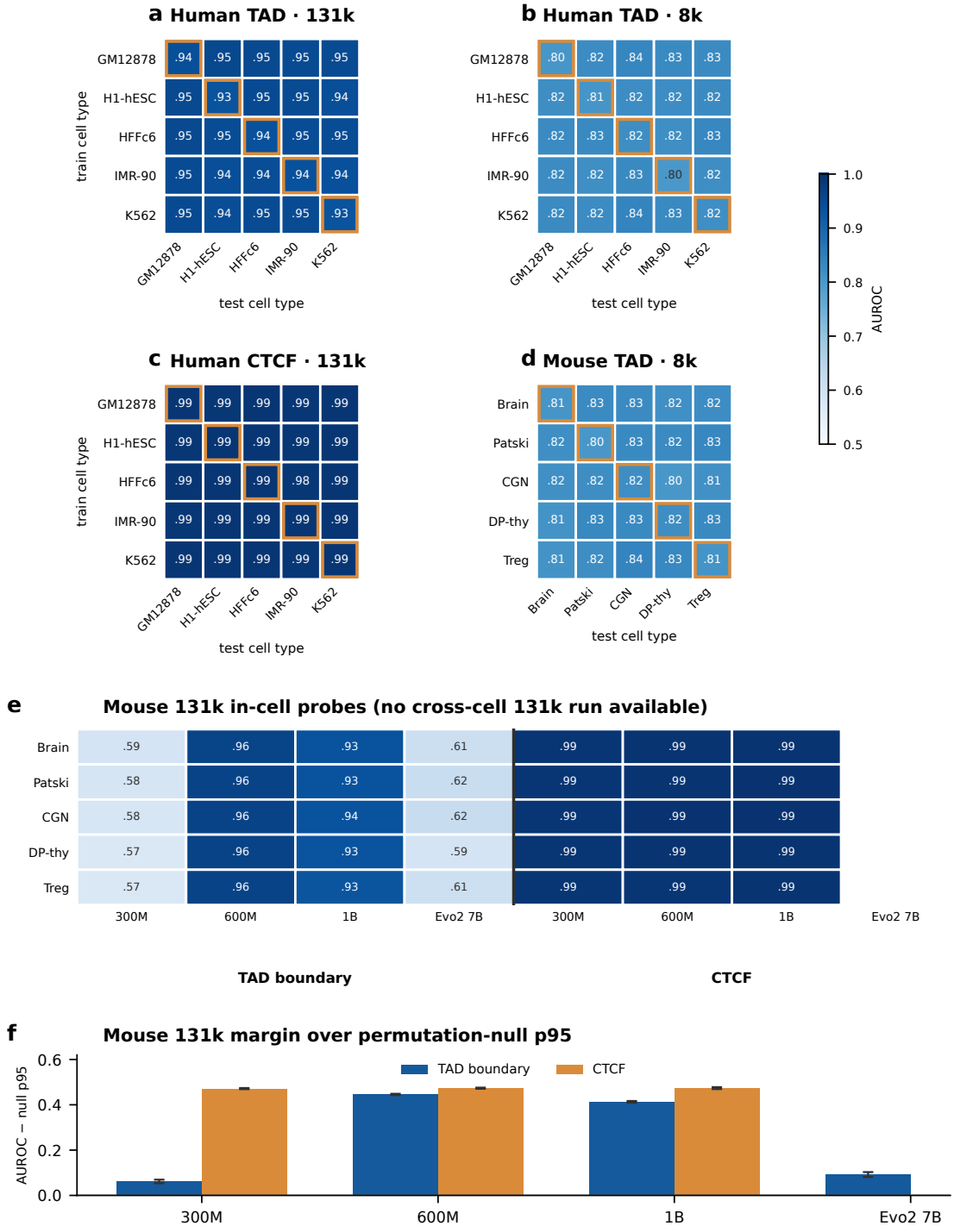

Supplementary Fig. S1 | **Per-stage training and validation loss for the continued-pretraining curriculum.** Training (solid) and validation (dashed) language-modelling loss versus stage-local tokens for the 600M and 1B CENO models during **a,** Stage II (8k context), **b,** Stage III (131k context) and **c,** Stage IV (1M context). Filled circles mark stage-end validation values (annotated). Curves are lightly denoised (robust outlier rejection followed by median/EWM smoothing); context length and data mixture differ across stages.



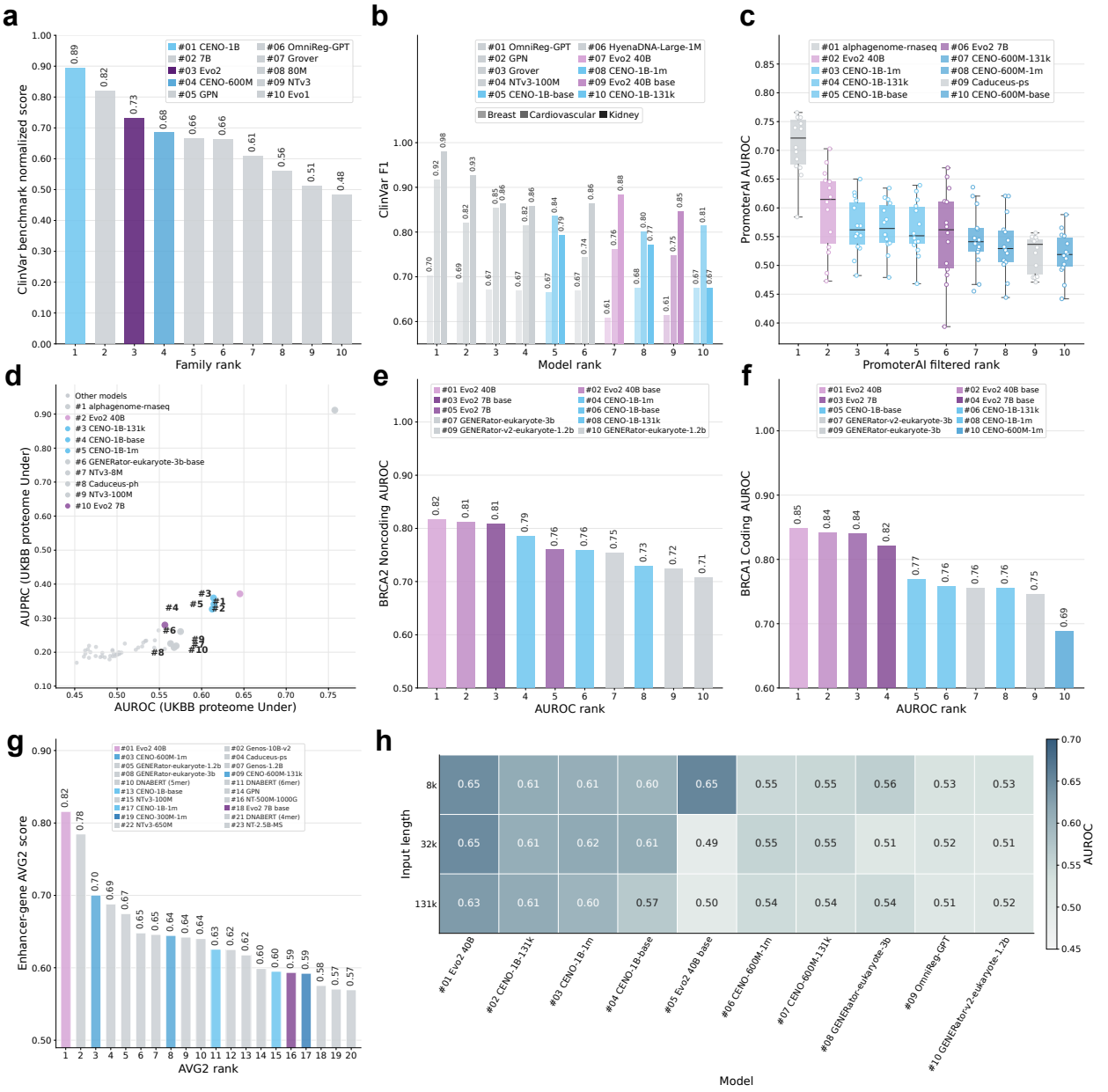

Supplementary Fig. S2 | **TAD-attention samples and cross-setting gains. a,** Same-locus 131k-minus-8k $\log_2$ within/across-TAD attention (W/A) gains for 24 retained matched TAD-boundary loci, ranked by CENO gain (rows labelled H, human; M, mouse). **b,** Per-locus $\log_2$ W/A for matched negative controls (left, grey) versus positive boundaries (right, coloured) in the 600M, 1B and Evo 2 7B 131k settings; connecting lines join the group means. **c,** Cross-setting 131k-minus-8k $\log_2$ W/A gains across five human and five mouse cell or tissue settings. Colour scale in **a** and **c** is shared.



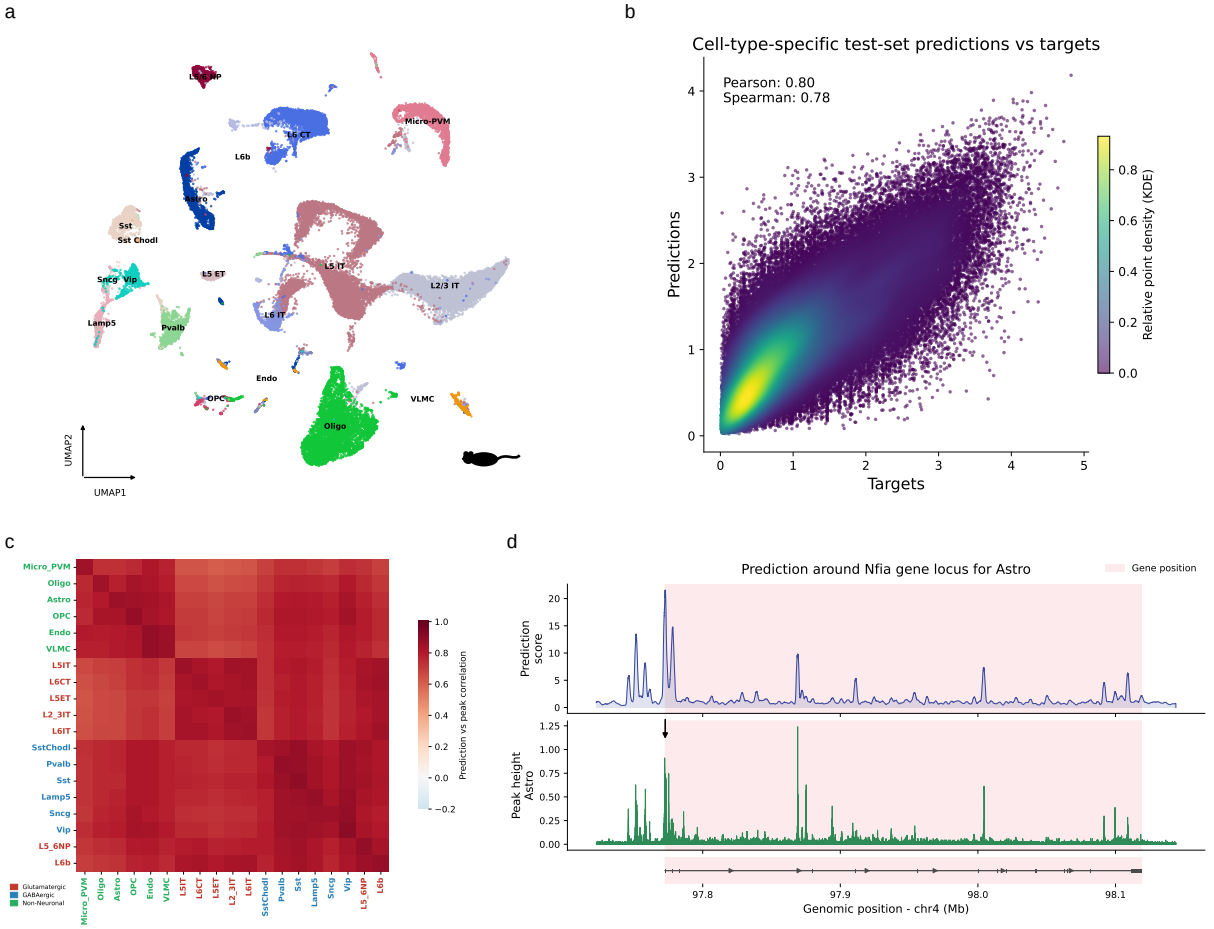

Supplementary Fig. S3 | **Annotation-linked attention: overall statistics and representative examples. a,** Fraction of bins exceeding one normalized attention unit for annotated (coloured) versus background (grey) bins across all successful cases in the boundary→promoter, boundary→CTCF and enhancer→promoter relationships; bars denote medians. **b,** Per-case comparison of the same fractions (background, $x$; annotated, $y$); all cases lie above the equality line, indicating a systematic rather than cherry-picked effect. **c–h,** Six representative CENO 1B 131k query-to-annotation profiles that are typical rather than top-ranked (three boundary→promoter, three boundary→CTCF; GM12878, K562, H1-hESC, HFFc6 and IMR-90). Traces show the normalized attention delta of the boundary query relative to non-annotation bins along the locus; shaded spans mark annotation bins, with the genomic coordinate track below. Insets give the mean normalized attention in annotated versus non-annotated bins (error bars, bootstrap 95% CI) and the percentage of bins above one $z$-unit.



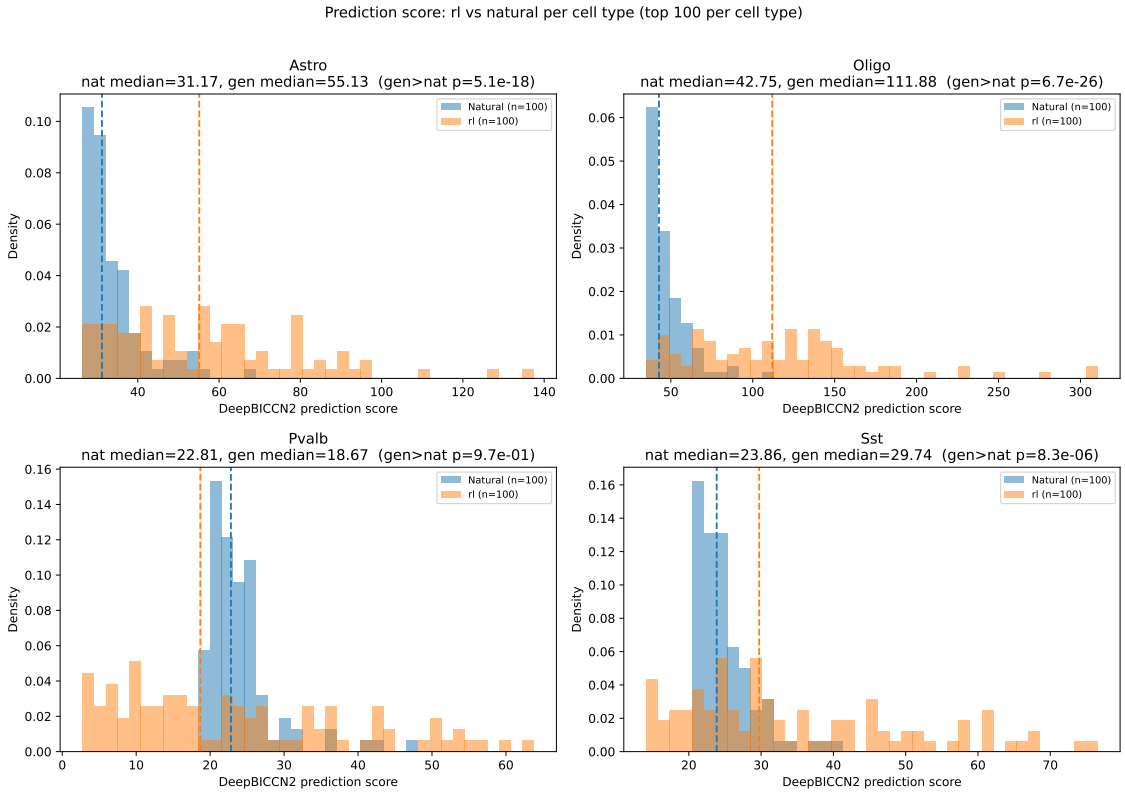

Supplementary Fig. S4 | **Frozen-state cross-cell-type TAD-boundary and CTCF probe transfer.** **a–d,** Train × test AUROC matrices for frozen-representation linear probes evaluated at the layer maximizing in-cell AUROC; orange boxes mark the in-cell diagonal (train=test). **a,** Human TAD boundary, 131k; **b,** human TAD boundary, 8k; **c,** human CTCF, 131k; **d,** mouse TAD boundary, 8k, the only context and probe with a mouse cross-cell-type run. **e,** Mouse 131k in-cell selected-layer AUROC for TAD-boundary and CTCF probes across cell or tissue settings and model scales; no mouse cross-cell-type run exists at 131k (Evo 2 7B CTCF not run). **f,** Mouse 131k selected-layer AUROC margin over the label-shuffle permutation-null p95 (mean ± s.d. across the five settings); all margins are positive.




Supplementary Fig. S5 | **Supplementary task-specific clinical, regulatory and BRCA benchmarks.** **a,** ClinVar family-level normalized score, with CENO model sizes shown separately. **b,** ClinVar F1 across clinical groups. **c,** PromoterAI filtered-rank summary across expression-perturbation subtasks. **d,** PromoterAI UKBB proteome under-expression AUROC–AUPRC frontier. **e,** BRCA2 noncoding SNV AUROC. **f,** BRCA1 coding SNV AUROC. **g,** Enhancer-gene AVG2 ranking. **h,** TraitGym AUROC by input length for the top-ranked models.


a
L6b_Vip
Micro-PVM
L6 CT
L6b
Astro
Sst
Sst Chodl
L5 ET
L5 IT
L2/3 IT
Sncg
Vip
Lamp5
Pvalb
L6 IT
Endo
OPC
VLMC
Oligo
UMAP2
UMAP1
b
Cell-type-specific test-set predictions vs targets
Pearson: 0.80
Spearman: 0.78
Predictions
Targets
Relative point density (KDE)
c
Micro_PVM
Oligo
Astro
OPC
Endo
VLMC
L5IT
L6CT
L5ET
L2_3IT
L6IT
SstChodl
Pvalb
Sst
Lamp5
Sncg
Vip
L5_6NP
L6b
Glutamatergic
GABAergic
Non-Neuronal
Prediction vs peak correlation
d
Prediction around Nfia gene locus for Astro
Gene position
Prediction score
Peak height Astro
Genomic position - chr4 (Mb)
e
center padding
left padding
full seq (2114): 0.856
0.820
0.636
0.763
0.595
Test Pearson correlation
Padded input length

Supplementary Fig. S6 | **Evaluation of the cell-type-specific accessibility oracle in mouse cortex. a,** UMAP visualization of mouse motor cortex cells from the BICCN multiome dataset, colored by subclass annotation. **b,** Density scatter plot comparing oracle predictions and observed chromatin accessibility on held-out cell-type-specific peaks. Each point represents one nonzero peak–cell type pair after log transformation, and color indicates relative point density estimated by KDE. **c,** Cross-cell-type correlation matrix between predicted and observed chromatin accessibility profiles across held-out consensus peaks. Rows and columns correspond to mouse cortex subclasses, with label colors indicating major cell groups. **d,** Locus-level oracle scoring at the astrocyte marker gene *Nfia*. The top track shows the predicted astrocyte accessibility profile generated by sliding-window scoring, and the bottom track shows the observed astrocyte ATAC signal from the matched BigWig track. The shaded region marks the *Nfia* gene body. **e,** Effect of padded input length and padding strategy on oracle performance. Bars show test Pearson correlation for center padding and left padding at 500 bp and 1,000 bp padded input lengths; the dashed line indicates the full 2,114 bp input setting.




Supplementary Fig. S7 | **Target-cell oracle score distributions of RL-generated and natural cell-type-specific CREs.** Histograms compare on-target oracle prediction scores between RL-generated sequences and natural cell-type-specific CREs for Astro, Oligo, Pvalb and Sst. For each cell type, the top 100 sequences ranked by target-cell score were used for both groups. Dashed lines indicate group medians. Statistical significance was assessed using a one-sided Mann–Whitney U test for higher scores in generated sequences.

## B. Supplementary Tables

Supplementary Table S1 | **Model configuration and pretraining recipe.** Block notation: M, Mamba-2 sequence-mixing block; E, MoE feed-forward block with 8 experts and top-2 routing; *, attention block; -, dense feed-forward block.

| Model | Total | Active | Layers | Hidden | Heads | Max context | Blocks | Layer pattern |
|---|---|---|---|---|---|---|---|---|
| 300M | 300M | 100M | 9 | 1,024 | 8 | 1,048,576 | M4 / E4 / *1 / -0 | `MEM*EMEME` |
| 600M | 600M | 200M | 20 | 1,024 | 16 | 1,048,576 | M9 / E8 / *2 / -1 | `M-MEM*EMEMEM*EMEMEME` |
| 1B | 1B | 400M | 38 | 1,024 | 16 | 1,048,576 | M17 / E17 / *4 / -0 | `MEMEM*EMEMEM*EMEMEMEMEM*EMEMEM*EMEMEME` |

| Model | Stage | Context | Tokens | Samples | GPUs | CP | Micro batch | Global batch | LR range |
|---|---|---|---|---|---|---|---|---|---|
| 300M | I | 8k | 1.0T | 122.1M | 64 | 1 | 32 | 1,024 | $1\times10^{-3}$–$1\times10^{-4}$ |
| 300M | II | 8k | 0.5T | 61.0M | 64 | 1 | 32 | 1,024 | $1\times10^{-4}$–$1\times10^{-5}$ |
| 300M | III | 131k | 500B | 3.8M | 32 | 1 | 1 | 64 | $2\times10^{-5}$–$2\times10^{-6}$ |
| 300M | IV | 1M | 500B | 0.5M | 32 | 4 | 1 | 8 | $5\times10^{-6}$–$5\times10^{-7}$ |
| 600M | I | 8k | 2.0T | 244.1M | 32 | 1 | 16 | 1,024 | $8\times10^{-4}$–$8\times10^{-5}$ |
| 600M | II | 8k | 1.3T | 158.7M | 32 | 1 | 16 | 1,024 | $8\times10^{-5}$–$8\times10^{-6}$ |
| 600M | III | 131k | 500B | 3.8M | 32 | 2 | 1 | 64 | $1.6\times10^{-5}$–$1.6\times10^{-6}$ |
| 600M | IV | 1M | 500B | 0.5M | 64 | 8 | 1 | 8 | $4\times10^{-6}$–$4\times10^{-7}$ |
| 1B | I | 8k | 3.5T | 427.2M | 64 | 1 | 8 | 1,024 | $5\times10^{-4}$–$5\times10^{-5}$ |
| 1B | II | 8k | 2.5T | 305.2M | 64 | 1 | 8 | 1,024 | $5\times10^{-5}$–$5\times10^{-6}$ |
| 1B | III | 131k | 500B | 3.8M | 32 | 4 | 1 | 64 | $1\times10^{-5}$–$1\times10^{-6}$ |
| 1B | IV | 1M | 500B | 0.5M | 64 | 16 | 1 | 8 | $2.5\times10^{-6}$–$2.5\times10^{-7}$ |

Supplementary Table S2 | **Data mixture weights.** Weights for Stage I, Stage II and the long-context continuation stages. Values are percentages of sampled training tokens.

| Data source | Stage I | Stage II | Stage III/IV |
|---|---|---|---|
| GTDB / IMG prokaryotic genomes | 22.00 | 8.00 | 18.18 |
| Metagenomes | 24.00 | 14.00 | 10.10 |
| Viral genomes | 3.00 | 4.00 | 3.03 |
| Organelle genomes | 2.00 | 2.00 | 1.01 |
| Eukaryotic genic windows | 25.00 | 35.00 | 16.16 |
| mRNA | 10.00 | 15.00 | 11.11 |
| mRNA splice/promoter windows | 8.00 | 14.00 | 11.11 |
| ncRNA | 3.00 | 5.00 | 4.04 |
| Promoter windows | 3.00 | 3.00 | 1.01 |
| NCBI eukaryotic genomes, Animalia | 0.00 | 0.00 | 16.16 |
| NCBI eukaryotic genomes, Plantae | 0.00 | 0.00 | 5.05 |
| NCBI eukaryotic genomes, Fungi | 0.00 | 0.00 | 2.02 |
| NCBI eukaryotic genomes, Protista | 0.00 | 0.00 | 0.51 |
| NCBI eukaryotic genomes, Chromista | 0.00 | 0.00 | 0.51 |
| Grouped prokaryotic/metagenomic/viral/organelle | 51.00 | 28.00 | 32.32 |
| Grouped eukaryotic/transcript/regulatory/complete-genome | 49.00 | 72.00 | 67.68 |

Supplementary Table S3 | **Needle retrieval summary.** For each model-stage and test-context setting, positive statistics summarize nine records (three insertion depths by three trials), whereas null statistics summarize six shuffled-needle controls (two insertion depths by three trials). Success denotes the fraction of positive trials exceeding the prespecified retrieval-score threshold of 0.8.

| Context | Model | Stage | Positive mean | Success | Null mean | Null max | Positive n | Null n |
|---|---|---|---|---|---|---|---|---|
| 32k | 300M | II | 4.438 | 1.00 | 0.113 | 0.146 | 9 | 6 |
| 32k | 300M | III | 4.536 | 1.00 | 0.364 | 0.612 | 9 | 6 |
| 32k | 300M | IV | 4.603 | 1.00 | 0.437 | 0.592 | 9 | 6 |
| 32k | 600M | II | 2.929 | 0.89 | 0.012 | 0.021 | 9 | 6 |
| 32k | 600M | III | 6.126 | 1.00 | 0.009 | 0.018 | 9 | 6 |
| 32k | 600M | IV | 6.072 | 1.00 | 0.008 | 0.015 | 9 | 6 |
| 32k | 1B | II | 4.220 | 1.00 | 0.015 | 0.022 | 9 | 6 |
| 32k | 1B | III | 6.333 | 1.00 | 0.012 | 0.025 | 9 | 6 |
| 32k | 1B | IV | 6.268 | 1.00 | 0.011 | 0.024 | 9 | 6 |
| 64k | 300M | II | 4.076 | 1.00 | 0.116 | 0.147 | 9 | 6 |
| 64k | 300M | III | 4.509 | 1.00 | 0.306 | 0.325 | 9 | 6 |
| 64k | 300M | IV | 4.636 | 1.00 | 0.380 | 0.472 | 9 | 6 |
| 64k | 600M | II | 1.213 | 0.33 | 0.012 | 0.017 | 9 | 6 |
| 64k | 600M | III | 6.181 | 1.00 | 0.009 | 0.023 | 9 | 6 |
| 64k | 600M | IV | 6.159 | 1.00 | 0.008 | 0.012 | 9 | 6 |
| 64k | 1B | II | 2.961 | 1.00 | 0.021 | 0.040 | 9 | 6 |
| 64k | 1B | III | 6.163 | 1.00 | 0.011 | 0.016 | 9 | 6 |
| 64k | 1B | IV | 6.122 | 1.00 | 0.009 | 0.013 | 9 | 6 |
| 131k | 300M | II | 3.689 | 1.00 | 0.114 | 0.135 | 9 | 6 |
| 131k | 300M | III | 4.070 | 1.00 | 0.341 | 0.435 | 9 | 6 |
| 131k | 300M | IV | 4.425 | 1.00 | 0.398 | 0.454 | 9 | 6 |
| 131k | 600M | II | 0.068 | 0.00 | 0.011 | 0.016 | 9 | 6 |
| 131k | 600M | III | 6.103 | 1.00 | 0.008 | 0.013 | 9 | 6 |
| 131k | 600M | IV | 5.972 | 1.00 | 0.007 | 0.012 | 9 | 6 |
| 131k | 1B | II | 1.611 | 0.89 | 0.020 | 0.031 | 9 | 6 |
| 131k | 1B | III | 6.067 | 1.00 | 0.011 | 0.018 | 9 | 6 |
| 131k | 1B | IV | 6.214 | 1.00 | 0.010 | 0.019 | 9 | 6 |
| 1M | 300M | II | 1.879 | 1.00 | 0.115 | 0.148 | 9 | 6 |
| 1M | 300M | III | 2.976 | 1.00 | 0.291 | 0.341 | 9 | 6 |
| 1M | 300M | IV | 4.049 | 1.00 | 0.319 | 0.373 | 9 | 6 |
| 1M | 600M | II | 0.007 | 0.00 | 0.011 | 0.017 | 9 | 6 |
| 1M | 600M | III | 0.861 | 0.33 | 0.007 | 0.010 | 9 | 6 |
| 1M | 600M | IV | 5.415 | 1.00 | 0.006 | 0.010 | 9 | 6 |
| 1M | 1B | II | 0.035 | 0.00 | 0.018 | 0.024 | 9 | 6 |
| 1M | 1B | III | 2.217 | 0.89 | 0.019 | 0.027 | 9 | 6 |
| 1M | 1B | IV | 5.713 | 1.00 | 0.009 | 0.011 | 9 | 6 |

Supplementary Table S4 | CENO post-training architecture schedules. Hybrid pattern symbols match Table S1. In the row-local/fusion schedule, L denotes a row-local layer that restricts context to the same MSA row, and F denotes a fusion layer that uses the packed causal context across rows.

| Model | Layers | Hidden size | Heads | Active params | Hybrid pattern | Row-local/fusion schedule |
|---|---|---|---|---|---|---|
| 300M | 9 | 1024 | 8 | 100M | `MEM*EMEME` | `LLFFFLLFF` |
| 600M | 20 | 1024 | 16 | 200M | `M-MEM*EMEMEM*EMEMEME` | `LLLLFFFLLLLFFFLLFFFF` |
| 1B | 38 | 1024 | 16 | 400M | `MEMEM*EMEMEM*EMEMEMEMEM*EME`<br>`MEM*EMEMEME` | `LLLLFFFLLLLFFFLLLLFFLLFFFLL`<br>`LLFFFLLLLFF` |

Supplementary Table S5 | **Cell-type-specific CRE supervised fine-tuning and reinforcement-learning hyperparameters.** Settings for the conditional CRE generator built on the 600M CENO backbone. The MinGap contrast is computed over the four design classes (Astro, Oligo, Pvalb, Sst).

| Stage | Setting | Value |
|---|---|---|
| Shared | Backbone | 600M CENO (Stage II) |
| Shared | Conditioning | single raw-text cell-type prefix token |
| Shared | Generation target length | 500 bp (center crop of 2,114 bp window) |
| Shared | Oracle input length | 2,114 bp (center-padded with `N`) |
| Shared | MinGap contrast classes | Astro, Oligo, Pvalb, Sst |
| SFT | MinGap filter | $\geq 5$ |
| SFT | Per-cell top-$K$ (train / val) | 8,000 / 1,000 |
| SFT | Loss | weighted autoregressive cross-entropy |
| SFT | MinGap weight $(m_0, s_0, w_{\min}, w_{\max})$ | $(10, 10, 0.5, 2.0)$ |
| SFT | Optimizer / schedule | AdamW / cosine |
| SFT | Learning rate | $5 \times 10^{-5}$ |
| SFT | Epochs | 6 |
| SFT | Augmentation | reverse complement |
| RL | Algorithm | GRPO (DAPO objective) |
| RL | Initialization | SFT checkpoint |
| RL | Reward | tail-only oracle MinGap, $\max(0, \cdot)$ |
| RL | Invalid-sequence penalty | applied |
| RL | Completions per prompt | 32 |
| RL | Sampling temperature | 0.9 |
| RL | KL coefficient | 0.03 |
| RL | Reward scaling | group-level |
| RL | Optimization steps | 300 |