SCALING LAWS AND ARCHITECTURAL FRONTIERS IN METAGENOMIC FOUNDATION MODELS

Geraldene Munsamy[1,*], Gavin Ayres[1], Jeremie Dona[1], Carla Greco[1], Daniel Anderson[1], Srijani Sridhar[1], William Chow[1], Aaron Kollasch[1], Robert Pecoraro[1], Tanggis Bohnuud[1], Keith Kam[1], Gus Minto-Cowcher[1], Marcus Leung[1], Hassan Sirelkhatim[2], John St. John[2], Ali Taghibakhshi[2], Tyler Shimko[2], Jared Wilbur[2], Timur Rvachov[2], Saee Paliwal[2], Eddie Calleja[2], Noelia Ferruz[3], Kevin K. Yang[4], Philipp Lorenz[1], Francesco Farina[1]

[1]Basecamp Research, [2]NVIDIA, [3]Centre for Genomic Regulation, Barcelona, [4]Microsoft Research

ABSTRACT

Foundation models for genomics have the potential to revolutionize therapeutic design, yet the optimal architectural choices for modeling the vast and diverse distribution of metagenomic data remain under-explored. In this work, we present the machine learning methodology behind EDEN, a family of metagenomic foundation models scaled up to 28 billion parameters and trained on 9.7 trillion nucleotide tokens. We provide a systematic empirical study of architectural trade-offs between autoregressive Transformers (Llama-style), State-Space Models (Mamba), and Long-convolutional architectures (Hyena) for nucleotide-level modeling. Contrary to recent trends favoring linear-time sequence models for long-range biological data, we demonstrate that the Llama architecture exhibits superior scaling efficiency and semantic retrieval capabilities as the model capacity grows. We derive a set of quality-aware scaling laws for metagenomics, showing how model performance follows predictable power-law behavior across three orders of magnitude in parameters and data. Through extensive benchmarking, spanning unsupervised zero-shot fitness prediction, semantic completion, and gene recovery, we establish a blueprint for scaling biological foundation models and provide empirical evidence demonstrating why Transformer-based architectures define the current frontier.

1 INTRODUCTION

The recent rise of biological language models has reframed biological sequences as a substrate for general-purpose representation learning and generative modeling, enabling transfer and controllable generation across tasks, scales, and organisms. These models, pretrained on raw sequences with self-supervised objectives, produce reusable representations and priors that can be adapted to data-scarce downstream settings. In genomics, this paradigm has already proved effective for learning transferable representations that support a range of downstream tasks, including regulatory element prediction (e.g., promoters, splice sites, and transcription factor binding sites) with pretrained DNA models such as DNABERT and its multi-species successor DNABERT-2 Ji et al. (2021); Zhou et al. (2024). Complementarily, large-scale genome-wide foundation models such as the Nucleotide Transformer further demonstrate transfer across diverse genomics benchmarks, including applications such as functional variant prioritization Dalla-Torre et al. (2025).

Metagenomics (Handelsman, 2004) studies the collective genetic material of microbial communities, yielding sequences from multiple taxa with highly uneven abundances that create a heterogeneous, long-tailed distribution challenging to model. At scale, metagenomic assemblies yield tens of thousands of contigs from species absent in reference databases (Parks et al., 2017; Nayfach et al., 2021), demonstrating that metagenomic sequence space far exceeds curated collections. Foundation models trained on this diversity have begun to demonstrate practical value in variant effect prediction, generation across biological modalities (Nguyen et al., 2024), and tracking viral evolution Zvyagin

*Corresponding author: geraldene@basecamp-research.com

et al. (2022), with applications in agriculture and human health (Liu et al., 2025; Munsamy et al., 2026). Metagenomic foundation models present a clear opportunity to act as general engines for (i) representation learning across microbial sequence space, (ii) scalable annotation of "microbial dark matter," and (iii) sequence generation to support downstream discovery pipelines.

However, the architectural and design choices that best support scaling in metagenomics remain under-explored. Recent work has proposed linear-time alternatives to attention, demonstrating million-token context at nucleotide resolution (Nguyen et al., 2023). Yet it remains unclear which model family provides the best trade-off between scaling efficiency, the ability to capture long-range dependencies, and practical training throughput at trillion-token regimes.

We address this gap with a controlled empirical study comparing three sequence-modeling families - autoregressive Transformers, state-space models, and long-convolutional architectures - trained across multiple scales on a large metagenomic corpus and evaluated on nucleotide-level likelihood and long-context behavior. Building on these results, EDEN (environmentally-derived evolutionary network) was developed (Munsamy et al., 2026), a family of metagenomic foundation models scaled up to 28 billion parameters and trained on up to 9.7 trillion nucleotide tokens from BaseData (Vince et al., 2025). The main contributions of this paper are:

- Systematic architectural comparison: We evaluate autoregressive Transformers (Llama), state-space models (Mamba), and long-convolutional architectures (Hyena) across three orders of magnitude in size (100M, 1B, 7B parameters). We find that while all architectures improve with scale, Transformers exhibit superior scaling efficiency on information-dense biological tasks.
- Scaling laws and data quality: We derive scaling laws for metagenomic language models, demonstrating predictable power-law behavior. Crucially, we quantify the impact of data quality: models trained on our curated, long-read-optimized dataset exhibit a significantly steeper scaling exponent than those trained on public short-read data, highlighting the critical role of assembly quality and diversity.
- Semantic vs. structural coherence: We uncover a fundamental trade-off in long-context generation. While linear-time models extrapolate structural properties (e.g., coding density) more gracefully beyond the training window, they generate sequences that diverge semantically from expected biological continuations. In contrast, Transformers maintain high semantic fidelity - correctly predicting downstream operon genes - demonstrating that global attention is essential for capturing the biological logic of gene regulation.

We emphasize that the primary goal of this work is not to provide an exhaustive benchmark of the final EDEN-28B model, but rather to detail the systematic scaling analysis and architectural insights that guided its development.

2 Related Work

DNA models. The development of genomics foundation models has broadly mirrored trends in NLP, moving from encoder-style pretraining toward large generative models. Early efforts such as DNABERT (Ji et al., 2021) adapted masked-language modeling to k-mer tokenized DNA and demonstrated transfer to discriminative genomics tasks including promoter, splice-site, and TF-binding prediction. DNABERT-2 (Zhou et al., 2024) extends this line of work to multi-species settings while replacing fixed k-mers with a learned BPE tokenizer and introducing a standardized benchmark for genome understanding. Along similar lines, the Nucleotide Transformer (Dalla-Torre et al., 2025) scales masked pretraining to models up to 2.5 billion parameters trained on broad genomic corpora, yielding transferable representations that support a wide range of downstream analyses, including variant-centric evaluations. In regulatory genomics, long-context sequence-to-function predictors such as Enformer (Avsec et al., 2021) and AlphaGenome (Avsec et al., 2025) have advanced functional track prediction over extended input windows.

Generative genomic architectures. A line of work explores generative architectures that learn nucleotide-level sequence priors for *de novo* generation and design. Evo-1 and Evo-2 (Nguyen et al., 2024; Brixi et al., 2025) demonstrate long-context genomic modeling and generation using the Striped Hyena family of hybrid convolutional architectures, emphasizing the role of scalable long-range

sequence operators. HyenaDNA (Nguyen et al., 2023) shows that long-convolutional models can reach megabase-scale context at single-nucleotide resolution with models up to 6.6M parameters. Beyond convolutions, state-space models have been adapted to genomics: the MambaDNA block in (Schiff et al., 2024) achieves strong performance on long-range variant-effect prediction benchmarks.

Scaling laws for language models. Natural language model performance follows predictable power-law scaling with compute, data, and parameters (Kaplan et al., 2020; Hoffmann et al., 2022). In biology, scaling laws have been studied for protein (Lin et al., 2023) and genomic models (Nguyen et al., 2024), though the latter was limited to 1B parameters.

Metagenomic corpora and foundation models. Recent work has explored foundation-model training directly on metagenomic data. While (Zvyagin et al., 2022) demonstrate the utility of language models for viral evolution, their design is heavily tailored to SARS-CoV-2, limiting generalization to broader biological classes. Closer to our approach, (Liu et al., 2024) adapt the Llama architecture for genomic data; however, their reliance on short-read sampling restricts their model to a narrow context window of 512bp. Recent work by (Zhou et al., 2025) has shown that scaling autoregressive architectures on metagenomic assemblies, from 100M to 4B parameters, yields significant gains in data representation quality and generation. Finally, (Nguyen et al., 2024) established a baseline scaling law comparing Transformer++, Mamba, and Hyena architectures. Crucially, however, their benchmarks were restricted to the 1B parameter regime, leaving the comparative behavior of these architectures at larger scales unexplored.

3 Data and experimental setup

3.1 Data sourcing and curation

Public genomic repositories exhibit significant taxonomic bias toward clinically relevant organisms (Hernandez et al., 2020), limiting model generalization (Ding & Steinhardt, 2024). To overcome these limitations, EDEN models are trained on BaseData (Vince et al., 2025), a corpus enriched for environmental and host-associated metagenomes comprising approximately 9.7 trillion nucleotide tokens (see Section B for details).

We apply stringent filtering criteria to ensure high data quality: contigs must exceed 2 kb in length, exhibit gene density $>$20%, and have sequencing depth $\geq$4X. Low-complexity and eukaryotic-viral sequences were excluded (see Section B for full filtering details).

3.2 Tokenization strategy

EDEN employs a byte-level tokenizer at single-nucleotide resolution with a vocabulary size of 512, accommodating canonical nucleotides (A, C, G, T) and special tokens.

3.3 Training pipeline

Genomic sequences were partitioned into overlapping windows of 8,192 tokens with 200 bp overlap to preserve local context at boundaries. Each window includes BOS, SEP, and EOS special tokens.

3.4 Model architectures and training

To inform the design of the EDEN-28B model and derive rigorous scaling laws, we evaluated three prevailing architectural paradigms across three orders of magnitude in size (100M, 1B, and 7B parameters):

- Llama3 (Transformer): An autoregressive baseline utilizing standard global attention ($O(N^2)$ complexity).
- Mamba2 (SSM): A selective State-Space Model offering linear-time scaling ($O(N)$) with sequence length.
- Hyena: A long-convolutional architecture optimized for long-context modeling.

The Mamba and Hyena architectures we train are hybrid architectures (Brixi et al., 2025) consisting of interleaved attention and Mamba or Hyena layers, respectively. Specifically, we employ the

Mamba-2-Hybrid and Striped Hyena configurations, where approximately one attention layer is interleaved for every seven recurrent/convolutional layers. This hybrid design reflects standard practice in the field, as pure linear-time models have been shown to underperform hybrids at scale (Waleffe et al., 2024). For brevity, we refer to these hybrid architectures as "Mamba" and "Hyena", highlighting the key architectural differentiator. All models were trained on identical subsets of our BaseData dataset or OpenGenome2 (OG2), for up to 350 billion tokens. We note that at equivalent parameter counts, architectures differ in FLOPs due to the computational cost of attention versus recurrent/convolutional operations (see Table 5); we discuss the implications of this in Section 4. We defer to Section C for further details on the training protocol.

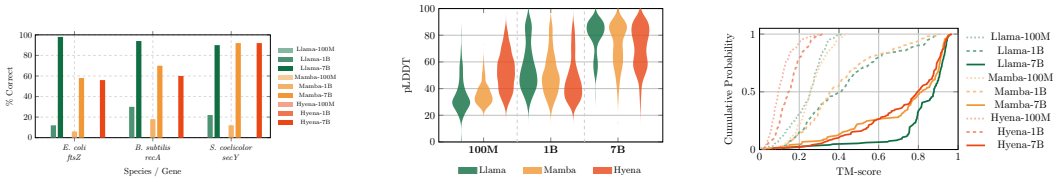
4 Scaling laws for metagenomics

Genomic modeling demands architectures that can efficiently resolve complex, non-local dependencies over long contexts. To identify the optimal architecture for this task, we evaluated the three candidate architectures described in Section 3 across three sizes (100M, 1B, and 7B parameters).

4.1 Benchmarking results: gene autocompletion

Gene autocompletion provides a stringent test of a genomic language model's ability to learn the "grammar" of DNA sequences. Unlike short-range prediction tasks, autocompletion probes whether the model can generate coherent, functionally plausible sequence continuations, reflecting an internalized representation of gene architecture and evolutionary conservation.

To evaluate autocompletion, we prompted each model with the first 20 % and 30% of a coding sequence and generated the remaining 80% and 70% autoregressively. We evaluated performance on three highly conserved housekeeping genes: *ftsZ* (cell division), *recA* (DNA repair), and *secY* (protein translocation), across three model organisms (*E. coli*, *B. subtilis*, *S. coelicolor*). Generated sequences were evaluated using Pfam annotations Paysan-Lafosse et al. (2025) and TM-scores (normalised to reference length) between ESMFold structures of proteins translated from the generations and wild-type proteins using USalign Zhang et al. (2022).



(a) Correct domain architecture (b) pLDDT score (c) TM-score

Figure 1: Structural quality metrics across architectures and scales. (a) Percentage of generated sequences with correct domain architecture. (b) Distribution of pLDDT scores. (c) Empirical cumulative distribution of TM-scores.

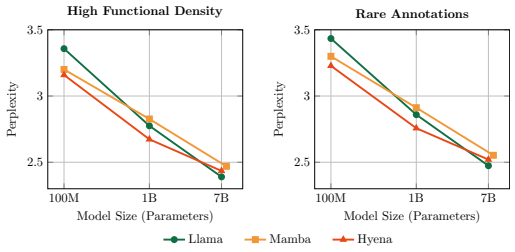
Across organisms and genes, smaller models showed limited generative capability often failing to recover the correct domain organization, whereas 7B-parameter models substantially improved domain-architecture recovery as seen in Figure 1a. This scaling effect is also reflected in structural predictions: Llama-based models produced the most reliable structures, with Llama-7B showing higher pLDDT values and closer agreement with expected folds in Figure 1b, consistent with a marked increase in predicted structural confidence relative to smaller models which most often yielded low-confidence structures. Furthermore, the TM-score distribution in Figure 1c is right-shifted for Llama, indicating closer agreement with reference folds and fewer failures (low TM-scores) compared to Mamba and Hyena equivalents. This trend holds across all three genes. In Section E, we provide an additional structural evaluation, including a breakdown of TM-scores by gene and the correlation between sequence identity and structural accuracy. Crucially, we observe that Llama maintains high structural fidelity even in regimes of low sequence identity. This indicates that the model captures biological semantics and structural constraints over mere sequence memorization.

Together, these results suggest that model capacity strongly influences biologically coherent gene completion and motivates our subsequent architecture selection for large-scale training.

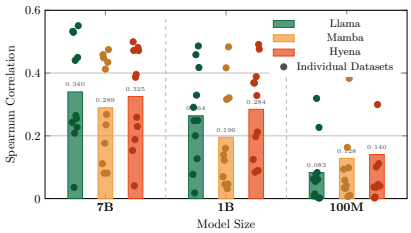
4.2 INFORMATION COMPRESSION VS. MODEL CAPACITY

Global metrics such as average perplexity are standard proxies for capacity in scaling law experiments, yet they often obscure local performance nuances. To better assess the semantic capacity of our models, we evaluate how each architecture copes with increasing functional density, using Kyoto Encyclopedia of Genes and Genomes (KEGG) annotations as a ground truth for biological signal (Kanehisa et al., 2016). We stratify the test set along two distinct axes:

- Functional Density (Complexity): We first measure how sequence-level perplexity evolves with the *count* of KEGG annotations present in a sequence. This tests how model design and scale influence the resolution of semantically rich, information-dense regions.
- Annotation Rarity (Generalization): We then measure how perplexity correlates with the *frequency* of specific KEGG annotations within the training corpus. This tests the model's ability to recall and generalize to rare or under-represented genetic elements.



(a) We evaluate architectural performance on (left) sequences with high functional density ($\geq$ 10 KEGG annotations) and (right) sequences containing rare annotations ($< 10^3$ occurrences in training).



(b) Zero-shot fitness prediction on DMS datasets. Bars represent the average Spearman correlation across all datasets, while individual points denote performance on specific DMS tasks

Figure 2: Semantic capacity and fitness prediction. (a) Information compression vs. model capacity across architectures and scales. (b) Zero-shot fitness prediction performance.

When analyzing scaling behavior through the prism of functional density in Figure 2a, we observe that Llama-based models exhibit superior scaling dynamics compared to their linear-time counterparts (Hyena and Mamba). While all architectures improve with size, Llama benefits most significantly from increased capacity, ultimately achieving best-in-class performance at the 7B parameter scale.

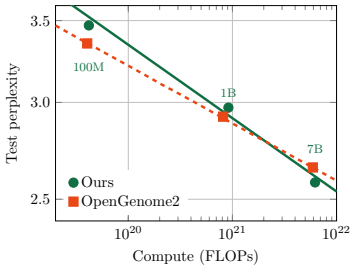
This architectural advantage is further supported by our generalization analysis. When stratifying sequences by the rarity of their KEGG annotations, we observe a distinct advantage for Llama on both rare and more common genetic elements, indicating that global attention is critical for resolving biological signals.

4.3 POWER-LAW BEHAVIOR AND DATA QUALITY

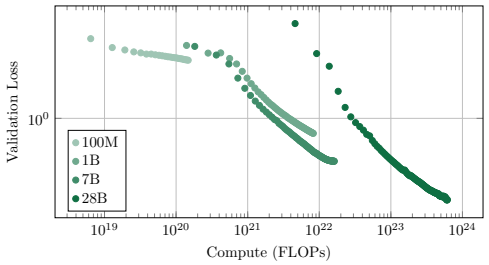
We observe that model performance follows a predictable power-law relationship between test loss (L) and compute (C), of the form $L(C) \propto C^{-\alpha}$. Following established methodology (Kaplan et al., 2020; Hoffmann et al., 2022), we fit scaling laws in log-log space using ordinary least squares regression on validation loss at convergence.

To isolate the impact of data quality on model performance, we conducted a controlled scaling analysis, benchmarking our BaseData corpus against the OpenGenome2 baseline Brixi et al. (2025). Crucially, we observe that the scaling exponent α is indeed data-dependent. Models trained on our in-house dataset follow a steeper power-law trajectory ($\alpha \approx 0.058$) compared to those trained on OpenGenome2 ($\alpha \approx 0.047$), see Figure 3a. This difference indicates that our dataset possesses a higher density of learnable signal, yielding superior performance gains for every unit of added model capacity.

We attribute this advantage to three key factors: (1) long-read sequencing yields longer contigs ($\approx$18 kb vs. $\approx$4 kb in OpenGenome2), enabling the model to resolve long-range dependencies such as operon architecture; (2) reduced fragmentation minimizes edge effects during tokenization, ensuring



(a) Models trained on our BaseData dataset (green, solid) exhibit a steeper scaling exponent compared to models trained on public OpenGenome2 data.



(b) Scaling behavior of EDEN models ranging from 100M to 28B parameters, validating the power-law extrapolation.

Figure 3: Scaling laws. (a) Data quality impact: our BaseData dataset yields steeper scaling compared to OpenGenome2. (b) Scale validation: EDEN-28B performance aligns with predicted loss.

functional units remain within the same context window; and (3) our sampling strategy mitigates bias toward reference genomes, providing richer signal from underrepresented taxa.

Statistical considerations. While fitting power laws to three data points limits statistical power, the fits exhibit high linearity ($R^2 > 0.985$ for both datasets). Loss trajectories in Section D provide additional evidence, showing clear performance separation at 7B scale.

Computational considerations. As shown in Table 5, architectures differ in FLOPs at equivalent parameter counts: at 7B, Mamba-Hybrid requires $\sim$2.2$\times$ the FLOPs of Llama, while Striped Hyena requires $\sim$0.9$\times$. Our primary comparison matches *parameter count* and *token budget*, which controls for capacity and data exposure.

4.4 Zero-shot fitness prediction

Foundation models have demonstrated strong zero-shot fitness prediction when sequence likelihood is used as a proxy for functional activity Bhatnagar et al. (2025); Nguyen et al. (2024); Dalla-Torre et al. (2025); Brixi et al. (2025). We evaluate this capability on the prokaryotic subset of RNAGym Arora et al. (2025), which contains deep mutational scans covering millions of variants. As illustrated in Figure 2b, Spearman correlation with measured fitness scales with model perplexity, with Llama excelling at the 7B scale.

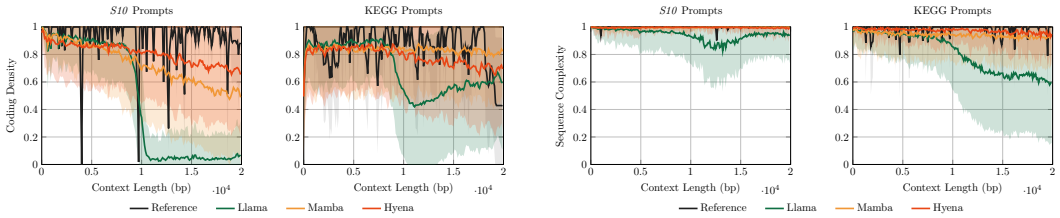
5 To the context length and beyond

5.1 Long-context generation

We employ a context window of 8,192 tokens, sufficient to capture bacterial operons (genes average $\sim$1000 base pairs (bp) Xu et al. (2006), operons $\sim$3 genes (Nuñez et al., 2013)) while enabling controlled cross-architecture comparison.

We evaluate the ability of different architectures to maintain biological fidelity over extended contexts. We prompt the models and allow them to generate continuously for up to 20,000 tokens. This rigorous setting tests the models' capacity to extrapolate well beyond their training context window and maintain structural coherence over long sequences.

We utilize two sets of prompts for evaluation: 1) The first set comprised ten sequence prompts, each corresponding to a distinct instance of the first gene of the ribosomal *S10* operon, where gene content and order is near universally-conserved across bacterial species (Table 6) (Coenye & Vandamme, 2005), and 2) representative sequences of the top seven most abundant genes in BaseData that have functional annotations derived from the KEGG database (Kanehisa & Goto, 2000). We measure coding density and sequence complexity (proportion of bases unmasked by DUST (Morgulis et al., 2006)) as a function of generated sequence length.



(a) Coding density.

(b) Sequence complexity.

Figure 4: Long-context generation metrics across context length for sequences generated with the *S10* and KEGG prompts. Shaded regions represent standard deviation.

Figures 4a and 4b illustrate coding density and sequence complexity, respectively. While all models start with high fidelity, their performance diverges as sequence length increases beyond the training context. The Llama model shows sharp degradation immediately after exceeding its training context of $\sim$8k tokens, whereas Mamba and Hyena exhibit more gradual decline, suggesting better extrapolation capabilities for linear-time architectures in the absence of explicit attention over the full history.

5.2 Semantic coherence

While coding density and sequence complexity provide proxies for structural validity, they, alone, do not measure biological correctness. To address this, we evaluate whether the models generate the genes expected to follow the prompt beyond the context window of 8,192 tokens (i.e., the downstream genes of the *S10* and *spc* operons, which together span a highly conserved region of approximately 10,600 bp). Figure 5 shows the frequency of expected gene recovery across generated sequences.

The *S10* and *spc* operon gene order are highly conserved across bacteria (Coenye & Vandamme, 2005), so a model that has learned bacterial genome structure should predict the correct downstream genes given the upstream context.

We observe a trade-off between structural stability and semantic coherence. Although Llama exhibits a sharper decline in coding density beyond its context window (Figure 4a), it recovers the expected downstream genes more often than the linear-time architectures, with Mamba coming close second (Figure 5). Notably, this is maintained beyond the 8,192 token context window used in training. Conversely, Hyena, while maintaining high sequence complexity and coding density, generates structurally valid sequences that semantically diverge from the expected *S10* and *spc* operon genes.

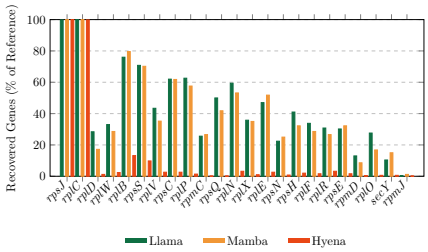


Figure 5: Percentage of generated contigs containing the expected genes for the *S10* benchmark.

6 Scaling to 28B parameters

Based on the analysis in Sections 4 and 5, we selected the Llama architecture for large-scale training. Despite the appeal of linear-time models for long sequences, Transformers consistently outperformed them on metrics critical for biological understanding: scaling efficiency, zero-shot fitness prediction, and semantic coherence.

The 28B model primarily validates our scaling law extrapolation rather than establishing state-of-the-art benchmarks; comprehensive downstream evaluation is beyond the scope of this methodology-focused paper.

6.1 EDEN-28B: VALIDATING THE LAW

Guided by the scaling laws established in Section 4, we trained the flagship EDEN-28B model on the full 9.7 trillion token corpus. As shown in Figure 3b, the final performance of EDEN-28B aligns closely with the predicted loss, validating the power-law extrapolation over orders of magnitude in scale. The trajectory from 100M to 28B parameters exhibits a remarkable consistency, with the validation loss decreasing predictably as a function of cumulative FLOPs. This trend extends beyond loss metrics: larger models generate genes with increasingly plausible structural and functional coherence. This smooth scaling behavior confirms that the model architecture effectively utilizes the increased compute budget and has not yet reached a point of diminishing returns given the vast size of the BaseData corpus. This predictability provides a robust blueprint for future scaling, suggesting that further gains can be achieved with even larger models and broader data distributions.

6.2 THE TRADE-OFF BETWEEN STRUCTURE AND SEMANTICS

Contrary to the prevailing trend towards linear-time models for long sequences, our empirical results at the trillion-token scale favor the Transformer architecture (see Table 1). We interpret this through the lens of the *semantic coherence* vs. *structural stability* trade-off observed in Sections 4 and 5.

Table 1: Performance per task across each 7B architecture.

Task	Llama	Mamba	Hyena
Mean pLDDT	**78.99**	72.65	70.82
Mean TM-score	**0.83**	0.73	0.73
Domain architecture recovery (%)	**94.00**	73.33	69.33
Mean perplexity across rare genes	**2.47**	2.55	2.52
Mean spearman correlation (RNAGym)	**0.339**	0.289	0.325
Long context - mean masking rates	15	4.2	**2.33**
Long context - mean coding density	0.6	0.78	**0.81**
Long context - mean number of ORFs with Pfams	**0.4**	0.36	0.2

While both Mamba and Hyena capture global structural properties such as coding density, they falter on tasks requiring precise semantics: Hyena struggles to reconstruct specific genetic identities, and Mamba underperforms on zero-shot fitness prediction. Linear-time architectures, despite incorporating selective attention, still compress or decay historical context due to their sub-quadratic designs. By contrast, Llama maintains lossless access to the full sequence history, enabling the precise retrieval needed for these information-dense tasks. As shown in Table 5, this comes without significant compute penalty: Llama's FLOPs cost is comparable to linear-time alternatives at equivalent parameter counts.

7 LIMITATIONS AND PRACTICAL CONSIDERATIONS

While EDEN demonstrates strong scaling and generation capabilities, several limitations warrant consideration. Transformers offer superior semantic accuracy but incur quadratic inference costs and limited structural extrapolation; linear-time models are preferable for ultra-long context or resource-constrained deployments despite weaker biological reasoning. Our model is trained solely on nucleotide sequences, potentially missing regulation encoded in epigenetic modifications or 3D chromatin structure. While our primary dataset is proprietary, we provide complete comparisons on the public OpenGenome2 dataset demonstrating reproducibility of key findings (see Section B).

8 CONCLUSION

In this work, we introduced the design and training methodology behind the EDEN family of foundation models for metagenomics. By rigorously benchmarking architectures, we demonstrated that Transformers offer superior scaling efficiency and semantic coherence compared to linear-time alternatives, while the latter excel at structural extrapolation - a trade-off practitioners should consider

based on their application requirements. We derived quality-aware scaling laws highlighting the critical role of data curation in model performance.

Crucially, evaluation of biological foundation models must evolve beyond perplexity, which is an imperfect proxy for biological plausibility. The true impact of models like EDEN will be determined by downstream tasks requiring synthesis of complex evolutionary rules - predicting variant fitness, designing functional pathways, or engineering microbial communities. The field must prioritize functionally-grounded benchmarks that distinguish causal biological understanding from pattern memorization.

REFERENCES

Rohit Arora, Murphy Angelo, Christian Andrew Choe, Courtney A. Shearer, Aaron W. Kollasch, Fiona Qu, Ruben Weitzman, Artem Gazizov, Sarah Gurev, Erik Xie, Debora S. Marks, and Pascal Notin. Rnagym: Large-scale benchmarks for rna fitness and structure prediction. bioRxiv, 2025. doi: 10.1101/2025.06.16.660049.

Žiga Avsec, Vikram Agarwal, Daniel Visentin, Joseph R. Ledsam, Agnieszka Grabska-Barwinska, Kyle R. Taylor, Yannis Assael, John Jumper, Pushmeet Kohli, and David R. Kelley. Effective gene expression prediction from sequence by integrating long-range interactions. Nature Methods, 18 (10):1196–1203, Oct 2021. ISSN 1548-7105. doi: 10.1038/s41592-021-01252-x.

Žiga Avsec, Natasha Latysheva, Jun Cheng, Guido Novati, Kyle R. Taylor, Tom Ward, Clare Bycroft, Lauren Nicolaisen, Eirini Arvaniti, Joshua Pan, Raina Thomas, Vincent Dutordoir, Matteo Perino, Soham De, Alexander Karollus, Adam Gayoso, Toby Sargeant, Anne Mottram, Lai Hong Wong, Pavol Drotár, Adam Kosiorek, Andrew Senior, Richard Tanburn, Taylor Applebaum, Souradeep Basu, Demis Hassabis, and Pushmeet Kohli. Alphagenome: advancing regulatory variant effect prediction with a unified dna sequence model. bioRxiv, 2025. doi: 10.1101/2025.06.25.661532.

Aadyot Bhatnagar, Sarthak Jain, Joel Beazer, Samuel C. Curran, Alexander M. Hoffnagle, Kyle S. Ching, Michael Martyn, Stephen Nayfach, Jeffrey A. Ruffolo, and Ali Madani. Scaling unlocks broader generation and deeper functional understanding of proteins. bioRxiv, 2025. doi: 10.1101/2025.04.15.649055.

Garyk Brixi, Matthew G. Durrant, Jerome Ku, Michael Poli, Greg Brockman, Daniel Chang, Gabriel A. Gonzalez, Samuel H. King, David B. Li, Aditi T. Merchant, Mohsen Naghipourfar, Eric Nguyen, Chiara Ricci-Tam, David W. Romero, Gwanggyu Sun, Ali Taghibakshi, Anton Vorontsov, Brandon Yang, Myra Deng, Liv Gorton, Nam Nguyen, Nicholas K. Wang, Etowah Adams, Stephen A. Baccus, Steven Dillmann, Stefano Ermon, Daniel Guo, Rajesh Ilango, Ken Janik, Amy X. Lu, Reshma Mehta, Mohammad R.K. Mofrad, Madelena Y. Ng, Jaspreet Pannu, Christopher Ré, Jonathan C. Schmok, John St. John, Jeremy Sullivan, Kevin Zhu, Greg Zynda, Daniel Balsam, Patrick Collison, Anthony B. Costa, Tina Hernandez-Boussard, Eric Ho, Ming-Yu Liu, Thomas McGrath, Kimberly Powell, Dave P. Burke, Hani Goodarzi, Patrick D. Hsu, and Brian L. Hie. Genome modeling and design across all domains of life with evo 2. bioRxiv, 2025. doi: 10.1101/2025.02.18.638918.

Tom Coenye and Peter Vandamme. Organisation of the s10, spc and alpha ribosomal protein gene clusters in prokaryotic genomes. FEMS Microbiology Letters, 242(1):117–126, 01 2005. ISSN 0378-1097. doi: 10.1016/j.femsle.2004.10.050.

Hugo Dalla-Torre, Liam Gonzalez, Javier Mendoza-Revilla, Nicolas Lopez Carranza, Adam Henryk Grzywaczewski, Francesco Oteri, Christian Dallago, Evan Trop, Bernardo P. de Almeida, Hassan Sirelkhatim, Guillaume Richard, Marcin Skwark, Karim Beguir, Marie Lopez, and Thomas Pierrot. Nucleotide transformer: building and evaluating robust foundation models for human genomics. Nature Methods, 22(2):287–297, Feb 2025. ISSN 1548-7105. doi: 10.1038/s41592-024-02523-z.

Frances Ding and Jacob Steinhardt. Protein language models are biased by unequal sequence sampling across the tree of life. In ICLR 2024 Workshop on Generative and Experimental Perspectives for Biomolecular Design, 2024.

Jo Handelsman. Metagenomics: application of genomics to uncultured microorganisms. Microbiol Mol Biol Rev, 68(4):669–685, December 2004.

Margarita Hernandez, Mary K Shenk, and George H Perry. Factors influencing taxonomic unevenness in scientific research: a mixed-methods case study of non-human primate genomic sequence data generation. Royal Society Open Science, 7(9):201206, 2020. doi: 10.1098/rsos.201206.

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, et al. Training compute-optimal large language models. In Proceedings of the 36th International Conference on Neural Information Processing Systems, pp. 30016–30030, 2022.

Yanrong Ji, Zhihan Zhou, Han Liu, and Ramana V Davuluri. Dnabert: pre-trained bidirectional encoder representations from transformers model for dna-language in genome. Bioinformatics, 37 (15):2112–2120, 02 2021. ISSN 1367-4803. doi: 10.1093/bioinformatics/btab083.

Minoru Kanehisa and Susumu Goto. Kegg: Kyoto encyclopedia of genes and genomes. Nucleic Acids Research, 28(1):27–30, 01 2000. ISSN 0305-1048. doi: 10.1093/nar/28.1.27.

Minoru Kanehisa, Yoko Sato, Masayuki Kawashima, Miho Furumichi, and Mao Tanabe. Kegg as a reference resource for gene and protein annotation. Nucleic acids research, 44(D1):D457–D462, 2016.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models. arXiv preprint arXiv:2001.08361, 2020.

Zeming Lin, Halil Akin, Roshan Rao, Brian Hie, Zhongkai Zhu, Wenting Lu, Nikita Smetanin, Robert Verkuil, Ori Kabeli, Yaniv Shmueli, et al. Evolutionary-scale prediction of atomic-level protein structure with a language model. Science, 379(6637):1123–1130, 2023.

Ollie Liu, Sami Jaghouar, Johannes Hagemann, Jeff Kaufman, and Willie Neiswanger. A foundation model for metagenomic sequences. In Neurips 2024 Workshop Foundation Models for Science: Progress, Opportunities, and Challenges, 2024.

Shaopeng Liu, Judith S. Rodriguez, Viorel Munteanu, Cynthia Ronkowski, Nitesh Kumar Sharma, Mohammed Alser, Francesco Andreace, Ran Blekhman, Dagmara Błaszczyk, Rayan Chikhi, Keith A. Crandall, Katja Della Libera, Dallace Francis, Alina Frolova, Abigail Shahar Gancz, Naomi E. Huntley, Pooja Jaiswal, Tomasz Kosciolek, Pawel P. Łabaj, Wojciech Łabaj, Tu Luan, Christopher Mason, Ahmed M. Moustafa, Harihara Subrahmaniam Muralidharan, Onur Mutlu, Nika Mansouri Ghiasi, Ali Rahnavard, Fengzhu Sun, Shuchang Tian, Braden T. Tierney, Emily Van Syoc, Riccardo Vicedomini, Joseph P. Zackular, Alex Zelikovsky, Kinga Zielińska, Erika Ganda, Emily R. Davenport, Mihai Pop, David Koslicki, and Serghei Mangul. Analysis of metagenomic data. Nature Reviews Methods Primers, 5(1):5, Jan 2025. ISSN 2662-8449. doi: 10.1038/s43586-024-00376-6.

Aleksandr Morgulis, E Michael Gertz, Alejandro A Schäffer, and Richa Agarwala. A fast and symmetric DUST implementation to mask low-complexity DNA sequences. J Comput Biol, 13(5): 1028–1040, June 2006.

Geraldene Munsamy, Gavin Ayres, Carla Greco, Keith Kam, Gus Minto-Cowcher, John St John, Tanggis Bohnuud, Matthew Bakalar, William Chow, Robert Pecoraro, Marcello der Torossian Torres, Aaron Kollasch, Marcus Leung, Hassan Sirelkhatim, Francesco Farina, Connor McGinnis, Srijani Sridhar, Daniel Anderson, Francesco Oteri, Ali Taghibakhshi, Jeremie Dona, Tyler Shimko, Cedric Stenbeeke, Alexandros Papadopoulos, Malcolm Krolick, Fabian Spoendlin, Purba Gupta, Sandeep Kumar, Anne Bara, Jared Wilbur, Noelia Ferruz, Timur Rvachov, Fangping Wang, Hanqun Cao, Hyun-Su Lee, Japan Mehta, Raphael Chaleil, Valerio Pereno, Sid Potti, Chris Emerson, Roy Tal Dew, Kevin K Yang, Eric Nguyen, Neha Tadimeti, Jillian F. Banfield, Alicia Frame, Emma Bolton, David Ruau, Rory Kelleher, Anthony Costa, Kimberley Powell, Cesar de la Fuente-Nunez, Glen-Oliver Gowers, Oliver Vince, Jonathan Finn, and Philipp Lorenz. Designing AI-programmable therapeutics with the EDEN family of foundation models. bioRxiv, 2026. doi: 10.64898/2026.01.12.699009.

Stephen Nayfach, Simon Roux, Rekha Seshadri, Daniel Udwary, Neha Varghese, Frederik Schulz, Dongying Wu, David Paez-Espino, I-Min Chen, Marcel Huntemann, Krishna Palaniappan, Joshua

Ladau, Supratim Mukherjee, T. B. K. Reddy, Torben Nielsen, Edward Kirton, José P. Faria, Janaka N. Edirisinghe, Christopher S. Henry, Sean P. Jungbluth, Dylan Chivian, Paramvir Dehal, Elisha M. Wood-Charlson, Adam P. Arkin, Susannah G. Tringe, Axel Visel, Helena Abreu, Silvia G. Acinas, Eric Allen, Michelle A. Allen, Lauren V. Alteio, Gary Andersen, Alexandre M. Anesio, Graeme Attwood, Viridiana Avila-Magaña, Yacine Badis, Jake Bailey, Brett Baker, Petr Baldrian, Hazel A. Barton, David A. C. Beck, Eric D. Becraft, Harry R. Beller, J. Michael Beman, Rizlan Bernier-Latmani, Timothy D. Berry, Anthony Bertagnolli, Stefan Bertilsson, Jennifer M. Bhatnagar, Jordan T. Bird, Jeffrey L. Blanchard, Sara E. Blumer-Schuette, Brendan Bohannan, Mikayla A. Borton, Allyson Brady, Susan H. Brawley, Juliet Brodie, Steven Brown, Jennifer R. Brum, Andreas Brune, Donald A. Bryant, Alison Buchan, Daniel H. Buckley, Joy Buongiorno, Hinsby Cadillo-Quiroz, Sean M. Caffrey, Ashley N. Campbell, Barbara Campbell, Stephanie Carr, JoLynn Carroll, S. Craig Cary, Anna M. Cates, Rose Ann Cattolico, Ricardo Cavicchioli, Ludmila Chistoserdova, Maureen L. Coleman, Philippe Constant, Jonathan M. Conway, Walter P. Mac Cormack, Sean Crowe, Byron Crump, Cameron Currie, Rebecca Daly, Kristen M. DeAngelis, Vincent Denef, Stuart E. Denman, Adey Desta, Hebe Dionisi, Jeremy Dodsworth, Nina Dombrowski, Timothy Donohue, Mark Dopson, Timothy Driscoll, Peter Dunfield, Christopher L. Dupont, Katherine A. Dynarski, Virginia Edgcomb, Elizabeth A. Edwards, Mostafa S. Elshahed, Israel Figueroa, Beverly Flood, Nathaniel Fortney, Caroline S. Fortunato, Christopher Francis, Claire M. M. Gachon, Sarahi L. Garcia, Maria C. Gazitua, Terry Gentry, Lena Gerwick, Javad Gharechahi, Peter Girguis, John Gladden, Mary Gradoville, Stephen E. Grasby, Kelly Gravuer, Christen L. Grettenberger, Robert J. Gruninger, Jiarong Guo, Mussie Y. Habteselassie, Steven J. Hallam, Roland Hatzenpichler, Bela Hausmann, Terry C. Hazen, Brian Hedlund, Cynthia Henny, Lydie Herfort, Maria Hernandez, Olivia S. Hershey, Matthias Hess, Emily B. Hollister, Laura A. Hug, Dana Hunt, Janet Jansson, Jessica Jarett, Vitaly V. Kadnikov, Charlene Kelly, Robert Kelly, William Kelly, Cheryl A. Kerfeld, Jeff Kimbrel, Jonathan L. Klassen, Konstantinos T. Konstantinidis, Laura L. Lee, Wen-Jun Li, Andrew J. Loder, Alexander Loy, Mariana Lozada, Barbara MacGregor, Cara Magnabosco, Aline Maria da Silva, R. Michael McKay, Katherine McMahon, Chris S. McSweeney, Mónica Medina, Laura Meredith, Jessica Mizzi, Thomas Mock, Lily Momper, Mary Ann Moran, Connor Morgan-Lang, Duane Moser, Gerard Muyzer, David Myrold, Maisie Nash, Camilla L. Nesbø, Anthony P. Neumann, Rebecca B. Neumann, Daniel Noguera, Trent Northen, Jeanette Norton, Brent Nowinski, Klaus Nüsslein, Michelle A. O'Malley, Rafael S. Oliveira, Valeria Maia de Oliveira, Tullis Onstott, Jay Osvatic, Yang Ouyang, Maria Pachiadaki, Jacob Parnell, Laila P. Partida-Martinez, Kabir G. Peay, Dale Pelletier, Xuefeng Peng, Michael Pester, Jennifer Pett-Ridge, Sari Peura, Petra Pjevac, Alvaro M. Plominsky, Anja Poehlein, Phillip B. Pope, Nikolai Ravin, Molly C. Redmond, Rebecca Reiss, Virginia Rich, Christian Rinke, Jorge L. Mazza Rodrigues, William Rodriguez-Reillo, Karen Rossmassler, Joshua Sackett, Ghasem Hosseini Salekdeh, Scott Saleska, Matthew Scarborough, Daniel Schachtman, Christopher W. Schadt, Matthew Schrenk, Alexander Sczyrba, Aditi Sengupta, Joao C. Setubal, Ashley Shade, Christine Sharp, David H. Sherman, Olga V. Shubenkova, Isabel Natalia Sierra-Garcia, Rachel Simister, Holly Simon, Sara Sjöling, Joan Slonczewski, Rafael Soares Correa de Souza, John R. Spear, James C. Stegen, Ramunas Stepanauskas, Frank Stewart, Garret Suen, Matthew Sullivan, Dawn Sumner, Brandon K. Swan, Wesley Swingley, Jonathan Tarn, Gordon T. Taylor, Hanno Teeling, Memory Tekere, Andreas Teske, Torsten Thomas, Cameron Thrash, James Tiedje, Claire S. Ting, Benjamin Tully, Gene Tyson, Osvlado Ulloa, David L. Valentine, Marc W. Van Goethem, Jean VanderGheynst, Tobin J. Verbeke, John Vollmers, Aurèle Vuillemin, Nicholas B. Waldo, David A. Walsh, Bart C. Weimer, Thea Whitman, Paul van der Wielen, Michael Wilkins, Timothy J. Williams, Ben Woodcroft, Jamie Woolet, Kelly Wrighton, Jun Ye, Erica B. Young, Noha H. Youssef, Feiqiao Brian Yu, Tamara I. Zemskaya, Ryan Ziels, Tanja Woyke, Nigel J. Mouncey, Natalia N. Ivanova, Nikos C. Kyrpides, Emiley A. Eloe-Fadrosh, and IMG/M Data Consortium. A genomic catalog of earth's microbiomes. Nature Biotechnology, 39(4):499–509, Apr 2021. ISSN 1546-1696. doi: 10.1038/s41587-020-0718-6.

Eric Nguyen, Michael Poli, Marjan Faizi, Armin Thomas, Michael Wornow, Callum Birch-Sykes, Stefano Massaroli, Aman Patel, Clayton Rabideau, Yoshua Bengio, Stefano Ermon, Christopher Ré, and Stephen Baccus. Hyenadna: Long-range genomic sequence modeling at single nucleotide resolution. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), Advances in Neural Information Processing Systems, volume 36, pp. 43177–43201. Curran Associates, Inc., 2023.

Eric Nguyen, Michael Poli, Matthew G. Durrant, Brian Kang, Dhruva Katrekar, David B. Li, Liam J. Bartie, Armin W. Thomas, Samuel H. King, Garyk Brixi, Jeremy Sullivan, Madelena Y. Ng, Ashley Lewis, Aaron Lou, Stefano Ermon, Stephen A. Baccus, Tina Hernandez-Boussard, Christopher Ré, Patrick D. Hsu, and Brian L. Hie. Sequence modeling and design from molecular to genome scale with evo. Science, 386(6723):eado9336, 2024. doi: 10.1126/science.ado9336.

Pablo A. Nuñez, Héctor Romero, Marisa D. Farber, and Eduardo P.C. Rocha. Natural selection for operons depends on genome size. Genome Biology and Evolution, 5(11):2242–2254, 11 2013. ISSN 1759-6653. doi: 10.1093/gbe/evt174.

Donovan H Parks, Christian Rinke, Maria Chuvochina, Pierre-Alain Chaumeil, Ben J Woodcroft, Paul N Evans, Philip Hugenholtz, and Gene W Tyson. Recovery of nearly 8,000 metagenome-assembled genomes substantially expands the tree of life. Nature microbiology, 2(11):1533–1542, 2017.

Typhaine Paysan-Lafosse, Antonina Andreeva, Matthias Blum, Sara Rocio Chuguransky, Tiago Grego, Beatriz Lazaro Pinto, Gustavo A Salazar, Maxwell L Bileschi, Felipe Llinares-López, Laetitia Meng-Papaxanthos, et al. The pfam protein families database: embracing ai/ml. Nucleic acids research, 53(D1):D523–D534, 2025.

Yair Schiff, Chia-Hsiang Kao, Aaron Gokaslan, Tri Dao, Albert Gu, and Volodymyr Kuleshov. Caduceus: Bi-directional equivariant long-range dna sequence modeling. arXiv preprint arXiv:2403.03234, 2024.

Oliver Vince, Phoebe Oldach, Valerio Pereno, Marcus H Y Leung, Carla Greco, Gus Minto-Cowcher, Saif Ur-Rehman, Keith Y K Kam, William Chow, Emma Bolton, Bupe R Mwambingu, Nadine Greenhalgh, Ineke Knot, Leif Christoffersen, Marlon Clark, Robert Pecoraro, Aaron W Kollasch, Tanggis Bohnuud, Matthew Bakalar, Philipp Lorenz, and Glen Gowers. Breaking through biology's data wall: Expanding the known tree of life by over 10x using a global biodiscovery pipeline. bioRxiv, 2025. doi: 10.1101/2025.06.11.658620.

Roger Waleffe, Wonmin Byeon, Duncan Riach, Brandon Norick, Vijay Korthikanti, Tri Dao, Albert Gu, Ali Hatamizadeh, Sudhakar Singh, Deepak Narayanan, et al. An empirical study of mamba-based language models. arXiv preprint arXiv:2406.07887, 2024.

Lin Xu, Hong Chen, Xiaohua Hu, Rongmei Zhang, Ze Zhang, and Z. W. Luo. Average gene length is highly conserved in prokaryotes and eukaryotes and diverges only between the two kingdoms. Molecular Biology and Evolution, 23(6):1107–1108, 04 2006. ISSN 0737-4038. doi: 10.1093/molbev/msk019.

Chengxin Zhang, Morgan Shine, Anna Marie Pyle, and Yang Zhang. Us-align: universal structure alignments of proteins, nucleic acids, and macromolecular complexes. Nature methods, 19(9): 1109–1115, 2022.

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

Zhihan Zhou, Robert Riley, Satria Kautsar, Weimin Wu, Rob Egan, Steven Hofmeyr, Shira Goldhaber-Gordon, Mutian Yu, Harrison Ho, Fengchen Liu, Feng Chen, Rachael Morgan-Kiss, Lizhen Shi, Han Liu, and Zhong Wang. Genomeocean: An efficient genome foundation model trained on large-scale metagenomic assemblies. bioRxiv, 2025. doi: 10.1101/2025.01.30.635558.

Maxim Zvyagin, Alexander Brace, Kyle Hippe, Yuntian Deng, Bin Zhang, Cindy Orozco Bohorquez, Austin Clyde, Bharat Kale, Danilo Perez-Rivera, Heng Ma, Carla M. Mann, Michael Irvin, J. Gregory Pauloski, Logan Ward, Valerie Hayot, Murali Emani, Sam Foreman, Zhen Xie, Diangen Lin, Maulik Shukla, Weili Nie, Josh Romero, Christian Dallago, Arash Vahdat, Chaowei Xiao, Thomas Gibbs, Ian Foster, James J. Davis, Michael E. Papka, Thomas Brettin, Rick Stevens, Anima Anandkumar, Venkatram Vishwanath, and Arvind Ramanathan. Genslms: Genome-scale language models reveal sars-cov-2 evolutionary dynamics. bioRxiv, 2022. doi: 10.1101/2022.10.10.511571.

A EDEN ARCHITECTURE DETAILS

All model architectures were implemented using the NVIDIA BioNeMo framework, which is built upon NeMo 2.0 and Megatron-LM. For the Llama architecture specifically, we utilized the Llama 3.1 implementation.

Table 2 provides the detailed configuration for the Llama models at different scales.

Table 2: Architecture configurations for Llama models at different scales.

Configuration	100M	1B	7B	28B
Layers	12	24	32	48
Hidden dimension	768	2048	4096	6144
Attention heads	12	16	32	48
KV heads (GQA)	12	16	8	8
FFN dimension	2048	5632	11008	16384
RoPE base freq	500000	500000	500000	500000

Table 3 provides the detailed configuration for the Mamba models at different scales Table 4 provides the detailed configuration for the Hyena models at different scales For the hybrid architectures (Mamba2-Hybrid and Striped Hyena), we follow the configurations from Brixi et al. (2025), where attention layers are interleaved with Mamba or Hyena layers at a ratio of approximately 1:7 (one attention layer per seven recurrent/convolutional layers).

B DATA

The metagenomic data used in this work was assembled while aiming at increasing taxonomic and geographic diversity relative to commonly used public corpora for training biological language models. The primary dataset collection and training data curation has been driven by consent considerations to ensure it is fair to stakeholders (public data largely ignores that) including an incentive structure through benefit sharing. Each of the 9.7 trillion tokens in pretraining of EDEN family can be traced back to consent of stakeholders.

Filtering criteria. We retain contigs (contiguous sequences of DNA assembled from many shorter reads) only if they exceeded 2 kb in length and exhibited a predicted gene density greater than 20%. To ensure high-confidence sequences, we required a minimum mean sequencing depth of 4X. Low-complexity contigs shorter than 10 kb and containing more than 50% low-complexity sequence, as quantified with DUSTmasker (v2.15.0), were removed from training Morgulis et al. (2006). Finally, for safety reasons, sequences with significant homology to known eukaryotic viruses were excluded to focus on prokaryotic and phage diversity.

Existing open datasets (e.g., OMG, OpenGenome, OpenGenome-2) often inherit similar limitations, including substantial redundancy and skewed representation across taxa and sampling locations. For example, sequences in major public repositories are heavily concentrated in a small number of species (e.g., in SRA, a large fraction of deposited sequence volume is attributable to a handful of

Table 3: Architecture configurations for Mamba models at different scales.

Configuration	100M	1B	7B
Layers	20	24	52
Hidden dimension	768	2560	4096
Attention heads	16	16	32
KV heads (GQA)	8	8	8
FFN dimension	4096	10240	21504
SSM groups	8	8	8
State dimensions	64	128	128
Head dimensions	96	160	64

Table 4: Architecture configurations for Hyena models at different scales.

Configuration	100M	1B	7B
Layers	11	25	32
Hidden dimension	768	1920	4096
Attention heads	24	15	32
FFN dimension	3072	5120	11008
Short filter	64	128	256
Medium filter	64	128	256
Long filter	768	1920	4096
RoPE base freq	10000	10000	10000

organisms). The dataset used here is designed to mitigate these biases by sampling broadly across underrepresented lineages. As an example, it includes sequences from 1M species-level groups that are not represented in standard reference collections. By reducing redundancy and taxonomic imbalance, we expect our analysis to provide more robust scaling-law estimates than prior studies that rely on more biased sequence collections.

C Training details and hyperparameters

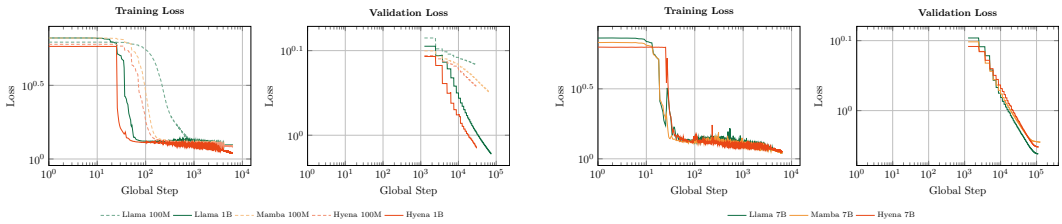
All models were trained using the AdamW optimizer with $\beta_1 = 0.9, \beta_2 = 0.95$, and a weight decay of 0.01. We used a cosine learning rate schedule with 2500 warm-up steps, reaching a peak learning rate of 3×10^{-5} for all models and decaying to 6×10^{-7} over 640920 steps. We stopped training once the target token count was reached, and used the resulting checkpoint for evaluation. Training was conducted on NVIDIA H200 GPUs using bfloat16 mixed precision with FP8 hybrid mode. To support model scaling, we utilized Tensor Parallelism (TP=2) and Sequence Parallelism across all runs. As detailed in the 100M and 1B models used a global batch size of 2048 distributed across 16 GPUs, while the 7B models utilized a batch size of 384 across 48 GPUs. The largest 28B model was scaled to 1008 GPUs across 126 nodes.

Table 5: Architecture scaling and training throughput summary

Architecture	#Params	#FLOPs	Global batch	#Eff. tok/batch	Total tokens processed
Mamba	100M	0.56×10^{21}	2048	16777216	356515840000
	1B	5.87×10^{21}	2048	16777216	356515840000
	7B	36.75×10^{21}	384	3145728	349965385728
Hyena	100M	0.28×10^{21}	2048	16777216	356515840000
	1B	2.64×10^{21}	2048	16777216	356515840000
	7B	14.33×10^{21}	384	3145728	349965385728
Llama	100M	0.11×10^{21}	2048	16777216	356515840000
	1B	2.32×10^{21}	2048	16777216	356515840000
	7B	16.36×10^{21}	384	3145728	349965385728

D Training and validation loss trajectories

We monitor the training stability and convergence through training and validation loss trajectories. Figure 6 shows the loss curves for the 100M/1B and 7B models, respectively. The 100M and 1B models (Figure 6a) were trained with a global batch size of 2048 sequences, while the 7B models (Figure 6b) used a batch size of 384 sequences. While all architectures exhibit consistent scaling behavior, we observe a shift in relative performance at scale: at the 7B parameter regime, the Llama model achieves the lowest validation loss, surpassing the recurrent architectures that performed comparably or better at smaller scales.



(a) 100M and 1B models

(b) 7B models

Figure 6: Training and Validation Loss Trajectories. (a) 100M and 1B models (dashed lines for 100M, solid for 1B). (b) 7B models.

E Additional benchmarking results

In addition to the aggregate TM-score metrics presented in the main text, we provide a more granular view of structural plausibility in Figure 7. This breakdown by gene and model highlights that while all architectures are capable of generating plausible structures, Llama-based models exhibit greater consistency, with fewer low-quality generations across all three genes.

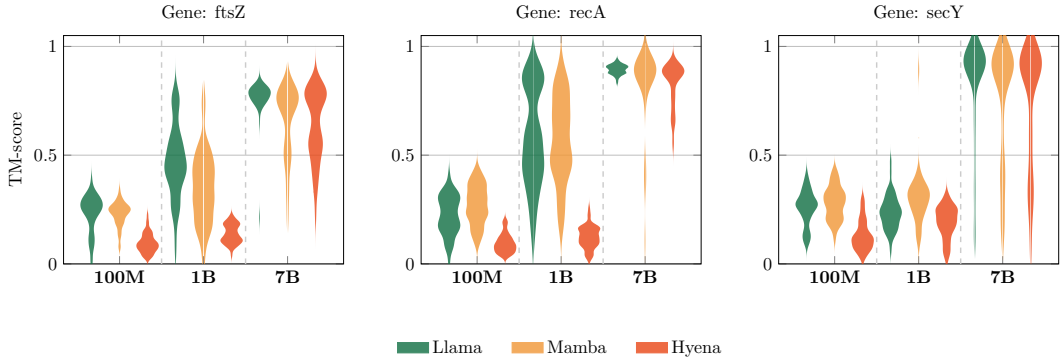


Figure 7: Distribution of TM-scores by gene and model. The violin plots highlight that while all models can generate plausible structures, Llama models exhibit greater stability with fewer low-quality generations.

Figure 8 shows the relationship between structural accuracy (TM-score) and sequence identity to the reference. Notably, Llama models maintain high TM-scores even at lower sequence identities, suggesting they generalize structural constraints better than Mamba and Hyena architectures.

F Long-context generation metrics

In this section, we provide a detailed breakdown of the long-context generation performance across different models. We evaluate the models on their ability to generate biologically plausible sequences that extend beyond the typical training context window, using a variety of metrics to assess sequence quality, including ORF length, coding density, GC content, and functional annotation rates.

Figure 9 presents a comprehensive comparison of these metrics for sequences generated by Reference, Llama, Mamba, and Hyena models. The results highlight the differing capabilities of each architecture in maintaining long-range coherence and biological fidelity. While all models capture basic sequence statistics, significant divergences appear in more complex metrics such as coding density and functional annotation rates (PFAM/KEGG), particularly for the Llama architecture which struggles to maintain these properties over extended generation lengths compared to the hybrid and recurrent architectures.

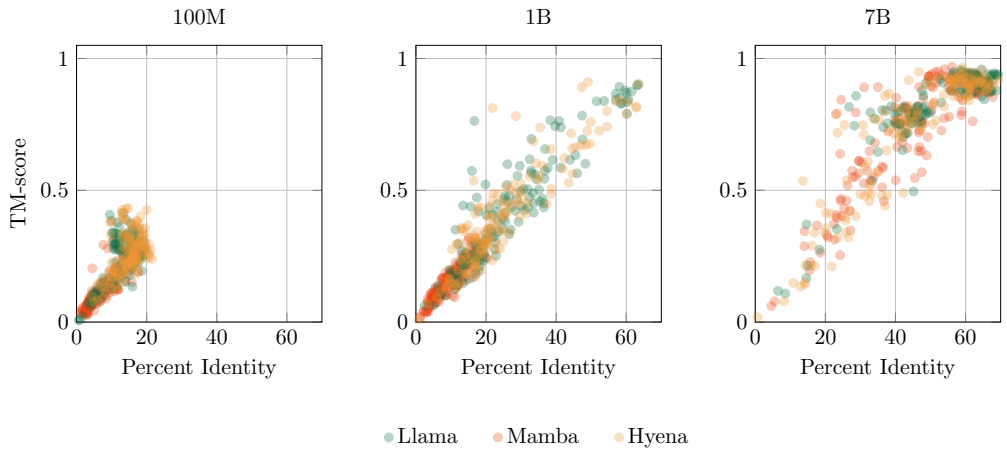


Figure 8: TM-score vs. Percent Identity. Larger models (7B) cluster in the upper-right quadrant. Llama models maintain high structural accuracy even at lower sequence identities, suggesting better generalization.

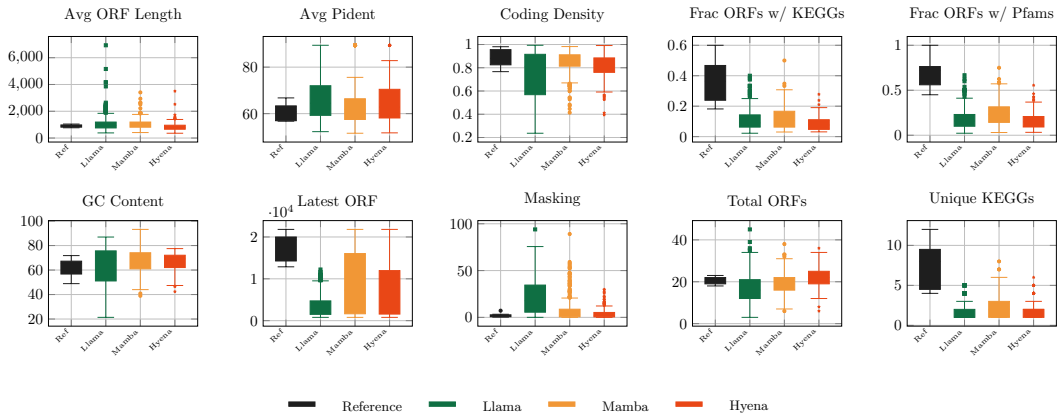


Figure 9: Detailed comparison of generation statistics for long-context KEGG prompts across Reference, Llama, Mamba, and Hyena models. The metrics include Average ORF Length, Average Percent Identity (Pident) to KEGG annotation, Coding Density, Fraction of ORFs with KEGG annotations/PFAM domains, GC Content, Latest ORF position, Masking rates, Total ORFs, and Unique KEGGs. Boxplots show the distribution of values for each metric, illustrating the variance and central tendency for each model.

Table 6: The accession numbers and coordinates of the *S10* prompts used in the long-context generation evaluations.

RefSeq accession	Genomic coordinates (start, end)	Orientation
CP026387	2875037, 2876036	Forward
CP028915	877336, 878335	Forward
NC_000913	3452271, 3453270	Reverse complement
NC_003197	3595538, 3596537	Reverse complement
NC_003198	4232972, 4233971	Forward
NC_004547	4502338, 4501337	Reverse complement
NC_009436	4054414, 4053413	Reverse complement
NC_009792	4350639, 4351638	Reverse complement
NC_013592	4267630, 4268629	Reverse complement
NC_017390	3682075, 3683074	Reverse complement

G IMPACT OF SCALE ON CONTEXT EXTENSION FOR LLAMA MODEL

The comparisons above utilize models trained on a controlled budget of 350 billion tokens to isolate architectural inductive biases. While the 7B Llama baseline exhibits degradation in structural metrics beyond its 8,192-token window, we find that scaling to 28B parameters offers negligible improvement in this regard. As illustrated in Figure 10, the EDEN-28B model, despite being trained on 9.7 trillion tokens, suffers from a similar degradation in both sequence complexity and coding density once generation exceeds the training context window. This observation suggests that the "context gap" is not merely a function of model scale or training duration but likely an intrinsic limitation of the architecture's ability to extrapolate beyond its positional training horizon.

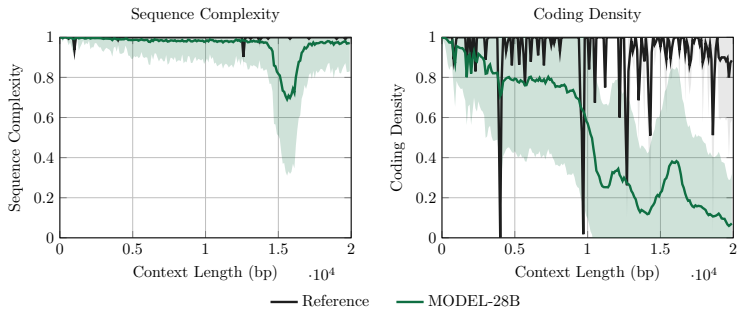


Figure 10: Sequence metrics for *S10* prompts. Comparison between the Reference (ground truth) and EDEN-28B model on (Left) Sequence Complexity and (Right) Coding Density. Despite the massive increase in scale, the EDEN-28B model fails to maintain metrics close to the reference distribution beyond the training context window ($\approx$ 8,192 bp), mirroring the limitations observed in smaller models.