Optimization and Tokenization Strategies for Biological Foundation Models: Evaluating H-Net and Muon

Anonymous Author(s)
Affiliation
Address
email

Abstract

Biological foundation models are powerful tools for modeling DNA and protein sequences, but their performance depends heavily on tokenization strategies—from BPE and k-mers to single-nucleotide resolution—each imposing rigid inductive biases. While recent architectures like Hyena and Mamba2 achieve strong performance using single nucleotide/amino acid resolution, this reliance on fixed granularity may not align with biology's natural organization. H-Net, a recently proposed architecture that replaces static tokenization with dynamic chunking learned end-to-end through gradient descent, offers a solution by allowing models to discover meaningful boundaries directly from biological data. We extend H-Net to biological sequences, incorporating a Projected Gated Convolutional (PGC) routing module to capture local motifs, and show that on parameter-matched HG38 pretraining H-Net outperforms Mamba2 while achieving strong performance on supervised protein tasks. We further evaluate the Muon optimizer, which has not previously been applied to proteins. Muon consistently improves convergence speed and stability across architectures, including H-Net, Mamba2, and Transformer, delivering both faster training and better final perplexity. These results highlight the value of exploring both architectural innovations and optimization methods as the field moves toward multimodal biological foundation models that require flexible and efficient training.

1 Introduction

The past few years have seen rapid progress in protein and DNA language models, with architectures such as Nucleotide Transformer NT, Evo2 Brixi et al. [2025] and ESM Lin et al. [2023] scaling to tens of billions of parameters. These models use self-supervised pre-training to learn rich representations of sequence, enabling downstream tasks including regulatory variant interpretation, enhancer–promoter modeling, and protein design. As biological foundation models grow in size, they provide a powerful framework for general-purpose genomic and proteomic modeling.

Existing foundation models rely on fixed tokenization schemes, character-level, BPE, or k-mers, that impose rigid sequence boundaries. This contrasts with biology, where short motifs combine into higher-order units such as regulatory elements or protein domains. Fixed-resolution tokenization forces models to reconstruct these patterns implicitly and often inefficiently. This challenge is magnified in multi-modal settings, where aligning heterogeneous inputs across resolutions is computationally costly without flexible, shared representations.

To address these limitations, we evaluate H-Net Hwang et al. [2025], a hierarchical architecture that replaces static tokenization with dynamic chunking learned end-to-end. H-Net adaptively segments sequences into variable-length chunks, compressing redundant regions while preserving informative

Submitted to 39th Conference on Neural Information Processing Systems (NeurIPS 2025). Do not distribute.

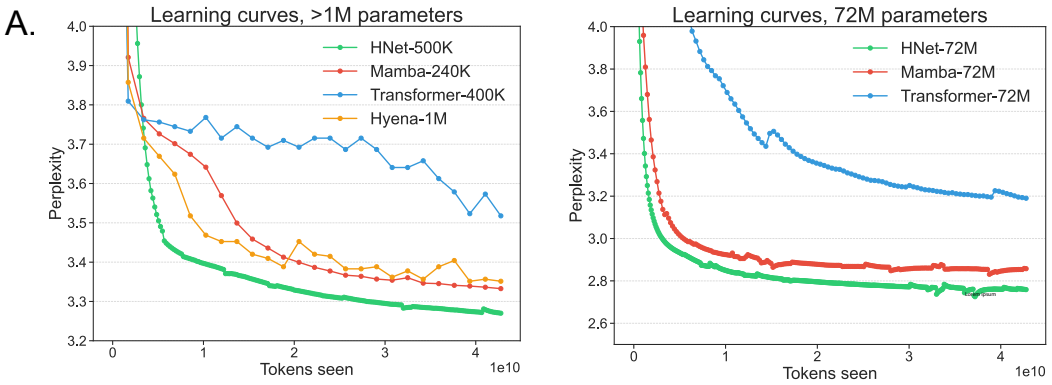

Figure 1: **(left)** Validation perplexity during training for sub-1M parameter models with an equivalent number of tokens seen **(right)** Validation perplexity during training for 72M parameter models with an equivalent number of tokens seen

features. Under compute-matched settings, H-Net reduces perplexity in DNA pretraining and achieves competitive performance on protein fitness benchmarks.

We also introduce Projected Gated Convolutions (PGC) between the embedding and routing modules. PGC layers combine depthwise convolutions, which capture local dependencies, with linear projections that gate the convolution and encode second-order interactions Ramesh et al.. This yields richer local features before chunking and is especially effective in protein modeling, where local motifs and domains are critical.

We evaluate the Muon optimizer muo across DNA and protein sequence modeling. While Muon has been applied to DNA, our work provides the most comprehensive assessment to date and the first application to proteins. Muon consistently improves convergence and stability across H-Net, Mamba, and Transformer, including on antibody and fluorescent protein DMS tasks.

In combination, H-Net and Muon provide complementary advances: H-Net offers a flexible alternative to fixed tokenization, while Muon accelerates convergence and improves stability across models. Together, these results demonstrate how tokenization strategies and optimizers can advance biological modeling.

1.1 Contributions

Our main contributions are as follows: **1. H-Net for biological sequences.** We apply H-Net to both DNA and protein language modeling, evaluating it under parameter- and data-matched settings. **2. Muon optimizer.** We provide the most comprehensive evaluation of the Muon optimizer to date on biological sequence modeling. Muon consistently improves convergence speed and stability for both pretraining and supervised tasks, across architectures including H-Net and Mamba. **3. Projected Gated Convolutions (PGC).** We introduce PGC modules into the H-Net pipeline. These combine depthwise convolutions with linear gating to capture second-order interactions and richer local dependencies, yielding gains in protein sequence modeling. **4. Protein tasks.** On supervised protein fitness/function benchmarks, H-Net with Muon is competitive with specialist models while using the same dynamic chunking interface.

2 Methods

Architecture

H-Net Hwang et al. [2025] is a hierarchical architecture that replaces static tokenization with learned dynamic chunking, allowing models to discover appropriate sequence granularities through gradient-based optimization. The model consists of encoder networks that process raw sequences, a routing module that determines chunk boundaries, a main network operating on compressed representations, and decoder networks that reconstruct full-resolution outputs. For biological sequences, we use

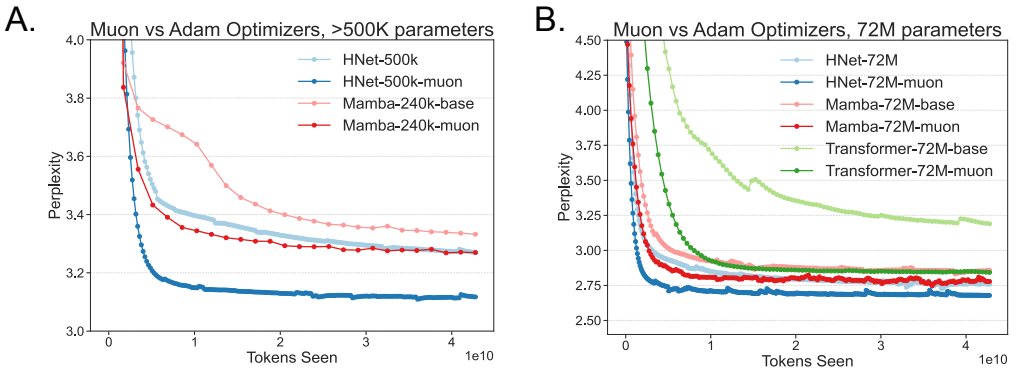

Figure 2: **(left)** Validation perplexity during training for sub-1M parameter models trained with and without Muon **(right)** Validation perplexity during training for 72M parameter models trained with and without Muon

single nucleotide/amino acid resolution as input, enabling the model to learn biologically relevant hierarchies without predetermined tokenization schemes.

Projected Gated Convolutional (PGC) Routing

We modify H-Net's original routing module to better capture local sequence dependencies. The original H-Net routing computes boundary probabilities using cosine similarity between adjacent positions, which we augment with convolutional features to capture local motifs.

Given encoder outputs $\hat{x}_t \in \mathbb{R}^D$ at position t, the PGC routing module computes:

$$q_t = W_q^{(1)} \hat{x}_t \odot \text{Conv1D}_q(\hat{x}_{t-k:t+k}) \tag{1}$$

$$k_t = W_k^{(1)} \hat{x}_{t-1} \odot \text{Conv1D}_k(\hat{x}_{t-1-k:t-1+k}) \tag{2}$$

where $W_q^{(1)}, W_k^{(1)} \in \mathbb{R}^{D \times D}$ are linear projection matrices, Conv1D_q and Conv1D_k are depthwise convolutional filters with kernel size 3 (padding 2), and $\odot$ denotes element-wise multiplication.

The boundary probability at position t is then computed using cosine similarity:

$$p_t = \frac{1}{2}\left(1 - \frac{q_t^\top k_t}{\|q_t\|\|k_t\|}\right) \in [0, 1], p_1 = 1 \tag{3}$$

Muon Optimizer

We employ the Muon optimizer, which achieves superior convergence through gradient orthogonalization via Newton-Schulz iteration. Unlike Adam which applies element-wise adaptive scaling, Muon orthogonalizes the momentum-averaged gradient to preserve directional information while removing magnitude imbalances. The update rule applies $W_{t+1} = W_t - \eta \times \sqrt{\text{fan-out}/\text{fan-in}} \times \text{NewtonSchulz}(M_t)$ where M_t is the momentum buffer. We use Muon with momentum coefficient 0.95 and learning rate 0.02 for matrix parameters, while applying AdamW to 1D parameters (biases, embeddings). This is the first demonstration of Muon's effectiveness on biological sequence tasks, particularly for supervised protein fitness prediction.

Pretraining on HG38

We pretrain parameter-matched models on human genome assembly HG38 using sequences of length 131,072 nucleotides. We compare four architectures: Mamba2, Hyena, Transformers, and H-Net, each trained with both AdamW and Muon optimizers. All models process single nucleotide resolution inputs and are trained for 10 epochs with batch size 256. Small and large model families are trained, ensuring fair evaluation of both architecture (dynamic chunking) and optimization (Muon).

Supervised Protein Fitness Tasks

We evaluate on RELSO protein fitness prediction tasks Castro et al. [2022], measuring validation Spearman correlation between predicted and experimental fitness values. The benchmark includes

Table 1: Evaluation of H-Net Ablations and Baselines on Protein Tasks (Spearman), with and without Muon. Best results per task are bolded.

	Mamba (750k)	Mamba (750k, Muon)	Hyena (750k)	Hyena (750k, Muon)	H-Net (750k)	H-Net (750k, Muon)	H-Net (750k, PGC)	H-Net (750k, PGC, Muon)
GFP	0.86	**0.86**	0.85	**0.86**	**0.86**	**0.86**	**0.86**	**0.86**
GB1_WU	**0.72**	**0.72**	0.71	0.71	0.70	**0.72**	**0.72**	**0.72**
Gifford	0.48	0.49	0.47	**0.50**	0.44	0.45	**0.50**	0.47

three diverse protein engineering datasets: GB1_WU (stability of protein G domain B1 variants), Gifford (CDR3 enrichment in antibody sequences), and GFP (green fluorescent protein brightness). These tasks require models to generalize from limited labeled examples to predict the functional impact of mutations. We train supervised model variants (Mamba2, Hyena, H-Net, H-Net with PGC, each with AdamW or Muon) for 250 epochs.

Evaluation For protein tasks, we report Spearman correlation on the RELSO validation sets. To assess learned chunking patterns, we visualize routing decisions on known functional elements and compute alignment with annotated protein domains and DNA regulatory regions.

3 Results

On HG38 DNA pretraining, H-Net consistently achieved the best perplexity among all architectures (Figure 1). At the sub-1M parameter scale, H-Net reached the lowest final perplexity (roughly 3.27), outperforming both Mamba2 (3.34) and Transformer (3.5). Scaling to 72M parameters reinforced this advantage, with H-Net converging to 2.76, compared to 2.85 and 3.18 for Transformers.

The Muon optimizer further boosted training efficiency and final quality across all architectures, accelerating convergence and lowering perplexity (Figure 2). At the small scale, Muon reduced perplexity for H-Net by 0.15, for Mamba2 by 0.06, and for Transformer by 0.08. At 72M parameters, Muon again improved every model: H-Net by 0.08, Mamba2 by 0.08, and Transformer by a striking 0.34. Among all experiments, H-Net with Muon achieved the best overall performance, reaching 2.68 perplexity at 72M parameters. Throughout all experiments, Muon paired improvements in perplexity with faster convergence.

For supervised protein fitness tasks, the results revealed task-specific patterns in architecture and optimizer effectiveness (Table 1). On GFP fluorescence prediction, most models achieved similar high performance (Spearman 0.85-0.86), suggesting this task may be approaching a performance ceiling with current approaches. GB1 WU stability prediction showed clearer differentiation, with Mamba2, H-Net with Muon, and H-Net-PGC variants all achieving the best performance (0.72), while standard H-Net lagged slightly (0.70-0.71). The Gifford CDR3 enrichment task proved most challenging and discriminative, with Hyena-Muon and H-Net-PGC achieving the highest correlation (0.50), followed by Mamba2-Muon (0.49), while standard H-Net variants performed poorly (0.44-0.47). Notably, while Muon consistently improved convergence during pretraining, its benefits for supervised tasks were more variable, suggesting that the optimizer's advantages may be most pronounced during large-scale self-supervised learning rather than fine-tuning on smaller labeled datasets.

4 Discussion

The Muon optimizer consistently improved convergence speed and stability across both DNA pretraining and supervised protein modeling. By reducing compute requirements while improving final performance, Muon highlights how optimization methods can deliver practical gains on par with architectural advances. These results suggest that the field should look beyond Adam and systematically explore optimizers and other ML training techniques, which may become increasingly important as biological foundation models scale.

H-Net provided a flexible alternative to fixed tokenization, with PGC further enriching local context in protein tasks. While improvements were more modest than Muon's, H-Net's adaptive chunking illustrates how tokenization strategies can shape downstream performance. Looking ahead, dynamic tokenization combined with robust optimization may be essential as models move toward multi-modal integration, where aligning heterogeneous data types under realistic compute budgets remains a core challenge.

References

Nucleotide Transformer: building and evaluating robust foundation models for human genomics | Nature Methods. URL https://www.nature.com/articles/s41592-024-02523-z.

Muon: An optimizer for hidden layers in neural networks | Keller Jordan blog. URL https://kellerjordan.github.io/posts/muon/.

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, February 2025. URL http://biorxiv.org/lookup/doi/10.1101/2025.02.18.638918.

Egbert Castro, Abhinav Godavarthi, Julian Rubinfien, Kevin B Givechian, Dhananjay Bhaskar, and Smita Krishnaswamy. Relso: a transformer-based model for latent space optimization and generation of proteins. *arXiv preprint arXiv:2201.09948*, 2022.

Sukjun Hwang, Brandon Wang, and Albert Gu. Dynamic Chunking for End-to-End Hierarchical Sequence Modeling, July 2025. URL http://arxiv.org/abs/2507.07955. arXiv:2507.07955 [cs].

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.

Krithik Ramesh, Sameed M Siddiqui, Albert Gu, Michael D Mitzenmacher, and Pardis C Sabeti. Lyra: An Efficient and Expressive Subquadratic Architecture for Modeling Biological Sequences.