HyProt: A Long-Context Hybrid Attention-SSM Protein Language Model and a Matched Comparison with Transformers

Lina Bellahmidi
InstaDeep
Paris, France

Ghassene Jebali*
InstaDeep
Paris, France

Arturo Fiorellini-Bernardis
InstaDeep
Paris, France

Sebastien Boyer
InstaDeep
Paris, France

Oliver Bent*
InstaDeep
Paris, France

Abstract

Protein language models are at one of two extremes in how they handle sequence context. Transformers such as ESM-2 attend to every pair of residues, which is precise but costs compute quadratic in sequence length. State-space models such as LC-PLM instead compress the sequence into a fixed-size memory, keeping the cost linear but giving up the exact recall of distant positions. Hybrid attention-SSM models occupy a useful middle ground for text and DNA. In this paper, we introduce HyProt, an 8M-parameter hybrid attention-SSM protein language model pretrained on UniRef50 with a masked language modeling objective and a 1024-token context, and compare it against a transformer trained with the same data, optimizer, and token budget, so that any difference comes from the architecture. HyProt performs on par with the matched transformer on ProteinGym. Evaluated out of distribution, on held-out sequences up to 16$\times$ the 1024-token training context, perplexity diverges at roughly 4K residues: below this the transformer is slightly ahead, while HyProt avoids the sharp degradation the transformer suffers above it. The throughput crossover falls around 8K residues. These results position the hybrid backbone as a practical middle point on the precision-efficiency spectrum for protein modeling. This is work in progress; we are scaling model size and adapting training for downstream tasks such as protein-protein interactions.

1 Introduction

Protein language models learn representations of protein sequences that transfer to tasks such as variant effect prediction, structure prediction, and interaction modeling (Rives et al., 2021), and most are transformers (Vaswani et al., 2017) trained with a masked language modeling objective (Devlin et al., 2019; Rives et al., 2021; Elnaggar et al., 2022; Lin et al., 2023). Self-attention compares every residue to every other, which lets contact maps and even atomic-level structure emerge from masked language modeling alone (Lin et al., 2023), at a cost quadratic in sequence length. ESM-2 therefore crops proteins to 1024 tokens, and although its rotary position embeddings (Su et al., 2021) allow extrapolation in principle, behavior past the training window is largely untested.

Length is not equally important everywhere. Most single proteins are short, so the quadratic cost does not limit them; it bites in the long tail of multi-domain proteins, and systematically whenever

*Corresponding authors: g.jebali@instadeep.com, o.bent@instadeep.com

40th Conference on Neural Information Processing Systems (NeurIPS 2026). Workshop: Representations for the Physical Sciences.

several chains are modeled together, as in protein complexes and protein-protein interactions (PPIs). MINT (Ullanat et al., 2026) applies self-attention within each chain and adds cross-chain blocks on top of ESM-2, partly to avoid treating a concatenation of chains as one sequence.

Hybrid models in language and DNA. Outside biology, the usual answer is not to replace attention but to mix it with something cheaper. Jamba (Lieber et al., 2024), Samba (Ren et al., 2024), Zamba (Glorioso et al., 2024), and Hymba (Dong et al., 2024) combine attention layers with Mamba-style state-space layers (Gu and Dao, 2023): attention gives precise recall within a bounded window, while the recurrent state carries compressed information over any distance at constant cost per token. Hymba also prepends learnable meta tokens acting as a small global memory (Dong et al., 2024), and HybriDNA applies the recipe to long DNA (Zhu et al., 2025).

Long-context protein models. For proteins, long-context work has focused on pure state-space models (SSMs): ProtMamba trains on concatenations of homologous sequences with per-sequence delimiters and position indices (Sgarbossa et al., 2025), and PTM-Mamba uses bidirectional gated Mamba blocks for post-translational modification prediction (Peng et al., 2024). Closest to this work, LC-PLM pretrains a bidirectional Mamba on UniRef50 with an MLM objective, reporting better scaling, length extrapolation, and gains over ESM-2 that are greatest for long proteins and complexes (Wang et al., 2024); PRIMA adds a two-phase schedule on concatenated interacting pairs (Fiorellini-Bernardis et al., 2026). The SSM end is viable; the middle of the spectrum has not been tested.

HyProt. We take the middle path. In HyProt the two mechanisms run side by side in the same block rather than one after the other: the attention branch uses a fixed window over nearby residues while the Mamba branch reads the whole sequence, so each block sees local detail and long-range context at once. Only the last layer uses global attention, so cost stays well below a transformer's as sequences get longer. We pretrain HyProt on UniRef50 with an MLM objective, compare it to a Protein Encoding Transformer (PET) trained under the same protocol at matched scale, and evaluate both on ProteinGym (Notin et al., 2023). Our contributions are:

1. **A hybrid attention-SSM protein language model.** Protein SSMs are pure-SSM and hybrid Mamba-transformers target text and DNA; we test hybrid models for proteins.
2. **A measurement of the mixed-memory trade-off.** We report how accuracy, compute, and memory change with protein sequence length against a matched transformer.
3. **Efficiency at scale.** We give throughput and peak-memory curves against context length, and show where the hybrid becomes faster than a transformer of the same size.

2 Methods

2.1 Dataset

We pretrain both models on UniRef50 (Suzek et al., 2015), the UniProt Reference Clusters at 50% identity, which keeps only one representative sequence per cluster and spans all domains of life. The corpus is therefore large and diverse while staying non-redundant, so the model cannot simply memorize near-identical sequences.

We use release 2026_01. We keep sequences of every length and do not filter short proteins. Sequences longer than 1024 amino acids are cropped to a random window of 1024 tokens, following Lin et al. (2023). We hold out 250,000 clusters for testing and train/validate on the rest. Sequences are tokenized with the ESM-2 alphabet (Lin et al., 2023). Batches are built with a token budget instead of a fixed number of sequences, and sequences of similar length are grouped together so that padding stays small. The same data, splits, cropping, and batching are used for both models.

2.2 Architecture

HyProt follows the hybrid-head design of Hymba (Dong et al., 2024). Each block passes one input to two branches at the same time: an attention branch with a sliding window of 512 tokens, so a residue only attends to its neighbors and cost grows linearly with length, and a bidirectional Mamba-2

branch (Dao and Gu, 2024) that reads the whole sequence both ways, carrying long-range context in a fixed-size state. The two outputs are combined and passed through a feed-forward layer. Only the last layer uses global attention; thanks to flash attention (Dao et al., 2022) the memory footprint stays linear in length at every layer, and with a single global layer the compute stays below a transformer's as we scale in length.

HyProt has 6 layers, hidden size 256, 8 attention heads, and 32 learnable meta tokens, trained embeddings that every residue can read (Dong et al., 2024), prepended to every input, shared across all proteins and read by both branches in every block, for 7.62M parameters in total. The attention branch uses rotary position embeddings (Su et al., 2021); the Mamba branch needs none, since its recurrence already depends on order. As a baseline, we train a PET with the ESM-2 architecture (Lin et al., 2023) at the same parameter count, using 6 layers, hidden size 320, 20 attention heads, full self-attention and rotary position embeddings.

2.3 Training

Both models are trained with the masked language modeling objective (Devlin et al., 2019). We select 15% of the tokens in each sequence; of these, 80% are replaced by a mask token, 10% by a random amino acid, and 10% are left unchanged. The loss is the cross-entropy between the predicted and the true token at the selected positions.

We train both models with AdamW (Kingma and Ba, 2015; Loshchilov and Hutter, 2019) for 800k steps at a batch of 524k tokens, 420B tokens in total. Appendix C gives the full hyperparameters. Training ran on 2 H100 GPUs and took 5.8 days for HyProt against 3.5 days for PET, so HyProt is about 1.7 times slower to train at this context length. This is expected: the crops are short enough that the quadratic term of attention is small, while the recurrent scan of the Mamba branch adds a fixed cost per token. The throughput advantage reported in Section 3.2.2 appears only once sequences get long. The two models share the same objective, optimizer, schedule, token budget, and hardware.

3 Results

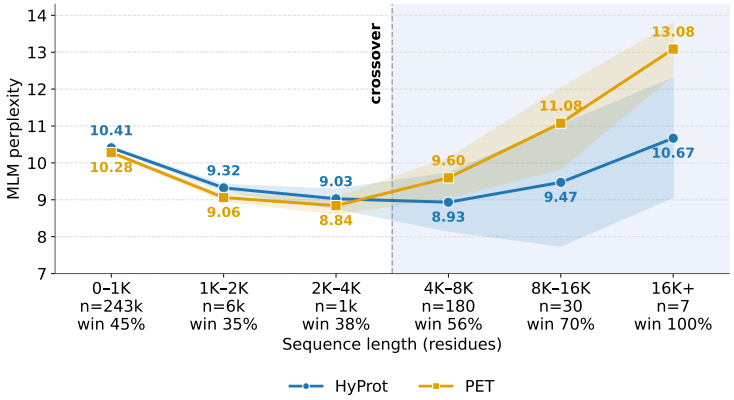
3.1 ProteinGym

We evaluate HyProt and PET zero-shot on ProteinGym (Notin et al., 2023), a large-scale benchmark of deep mutational scanning (DMS) assays measuring the fitness of protein variants relative to a wild type. We use the v1.3 substitution benchmark: 217 assays spanning 2.5M single and multi-point mutants, grouped under five functional categories: activity, binding, expression, organismal fitness, and stability. Adapting the scoring pipeline from the official ProteinGym ESM baseline (Notin et al., 2023), each mutant is scored from the model's masked marginal likelihood, and we report the Spearman correlation between predicted and measured fitness, averaged within each category. SE is the non-parametric bootstrap Standard Error over $10{,}000$ assay resamples and applies to the Overall column only; differences smaller than it are within sampling noise.

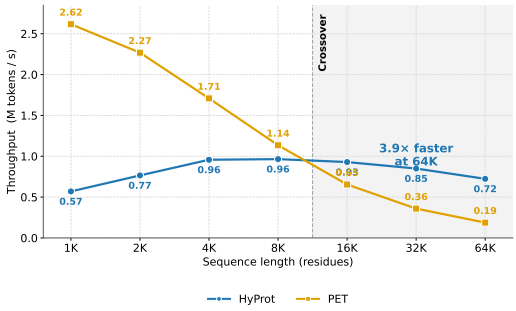
Table 1: Zero-shot ProteinGym Spearman ρ by functional category.

Model	Overall (217)	SE	Binding (n=13)	Expression (n=18)	Stability (n=66)	Activity (n=43)	Org. Fitness (n=77)
HyProt	0.2554	± 0.014	0.2707	0.2790	0.3231	0.2419	0.1545
PET	0.2553	± 0.014	0.2582	0.2846	0.3233	0.2439	0.1566

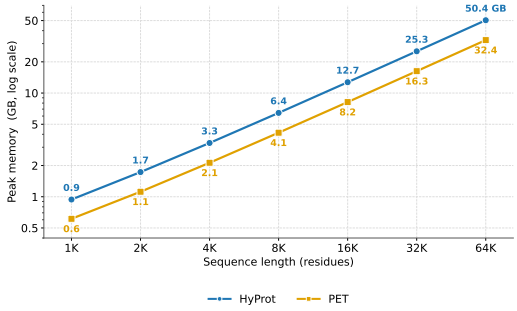
HyProt matches PET's fitness prediction at equal parameter budget; the headline metric is a tie. The largest margin is in the binding category, though with only 13 assays this is preliminary. It is still an encouraging signal for downstream binding tasks such as protein-protein interaction. PET recovers marginal leads on the rest of the categories, but Organismal fitness seems to be the hardest category for both models.



(a)



(b)



(c)

Figure 1: (a) Zero-shot MLM perplexity on the test set by sequence-length interval, with 95% CI. (b) Long-context inference throughput. (c) Peak inference memory. Lower is better for perplexity and memory; higher is better for throughput.

3.2 Long-context performance

3.2.1 Zero-shot perplexity stratification

After training, we measure how well each model fits held-out sequences as a function of length. We report token weighted masked language modeling perplexity on a held-out UniRef50 test split (250K sequences), masking 15% of residues under identical masks for both models. Sequences are fed at full length up to 16,384 residues, well beyond the 1024 residue training context, and grouped into fixed length ranges (Figure 1a); the longest bins hold few sequences, hence their wide confidence intervals.

PET is slightly better than HyProt up to $\sim$4K residues, where the curves cross. Both therefore extrapolate well to roughly four times their 1024-token training context. Beyond that PET degrades sharply while HyProt degrades far more gradually, which represents the language modeling counterpart to the throughput advantage that emerges with some length offset (Section 3.2.2).

3.2.2 Inference efficiency

We benchmark single forward inference on one H100-80GB across a context length sweep, batch 32, measuring throughput and peak memory (Figure 1b, 1c). At each length L, we feed a batch of 32 sequences of exactly L random residues with no padding. HyProt degrades gracefully when sequence length increases, whereas PET's $O(L^2)$ attention collapses. The curves cross slightly above 8K tokens, beyond which HyProt is strictly faster, reaching $3.9\times$ PET's throughput at 64K residues. Because PET uses flash attention, peak memory is linear in L for both models. HyProt's is consistently about $1.5\times$ higher, since its two branches run in parallel rather than in sequence,

every block pays for both the attention and the bidirectional scan. The efficiency gain is therefore a compute/throughput win at a constant extra 50% memory cost, rather than a memory win.

Single proteins rarely reach the crossover lengths, so the gain does not apply to them. It applies to the setting we are targeting next: complexes and multi-chain inputs, where several chains are joined into one long input and total length grows with the number of chains.

4 Conclusion and Outlook

We trained a parallel hybrid attention-SSM protein language model (HyProt) and compared it to a transformer (PET) trained the same way. At 8M parameters, the two are indistinguishable on ProteinGym overall within bootstrap error, with a non-significant lead on binding that motivates follow-up on interaction tasks. The difference appears with length: on held-out MLM perplexity, PET is slightly ahead up to 4K residues, beyond which it degrades sharply while HyProt degrades far more gradually. HyProt is also faster beyond $\sim$8K, reaching $3.9\times$ PET's throughput at 64K, at a constant extra 50% memory cost. So past the area of 4K to 8K residues, HyProt shows both better MLM perplexity and higher inference throughput than the matched transformer.

The small model size may limit the ability to exploit additional pretraining data. The same effect is reported for ESM-2 at this scale (Lin et al., 2023). The next step is to train larger models and see how the gap between HyProt and PET evolves with scale.

We also plan to post-train HyProt for protein-protein interaction. Two or more chains are joined into one sequence with separator tokens, and the model reads the whole set of chains as a single input. This is the case where the length advantage matters most, because complexes are longer than single sequences.

References

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 (ICML)*, 2024. URL `https://arxiv.org/abs/2405.21060`.

Tri Dao, Daniel Y. Fu, Stefano Ermon, Atri Rudra, and Christopher Ré. FlashAttention: Fast and memory-efficient exact attention with IO-awareness. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 35, 2022. URL `https://arxiv.org/abs/2205.14135`.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*, pages 4171–4186, 2019.

Xin Dong, Yonggan Fu, Shizhe Diao, Wonmin Byeon, Zijia Chen, Ameya Sunil Mahabaleshwarkar, Shih-Yang Liu, Matthijs Van Keirsbilck, Min-Hung Chen, Yoshi Suhara, Yingyan Lin, Jan Kautz, and Pavlo Molchanov. Hymba: A hybrid-head architecture for small language models. *arXiv preprint arXiv:2411.13676*, 2024. URL `https://arxiv.org/abs/2411.13676`.

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

Arturo Fiorellini-Bernardis, Sebastien Boyer, Jean Quentin, Ghassene Jebali, and Oliver Bent. PRIMA: A bidirectional state-space architecture and training approach for sequence modelling of protein-protein interactions. In *3rd Workshop on Multi-modal Foundation Models and Large Language Models for Life Sciences, International Conference on Machine Learning (ICML)*, 2026. URL `https://icml.cc/virtual/2026/72111`.

Paolo Glorioso, Quentin Anthony, Yury Tokpanov, James Whittington, Jonathan Pilault, Adam Ibrahim, and Beren Millidge. Zamba: A compact 7b ssm hybrid model. *arXiv preprint arXiv:2405.16712*, 2024. URL `https://arxiv.org/abs/2405.16712`.

Albert Gu and Tri Dao. Mamba: Linear-time sequence modeling with selective state spaces. *arXiv preprint arXiv:2312.00752*, 2023. URL `https://arxiv.org/abs/2312.00752`.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations*, 2015.

Opher Lieber, Barak Lenz, Hofit Bata, Gal Cohen, Jhonathan Osin, Itay Dalmedigos, Erez Safahi, Shaked Meirom, Yonatan Belinkov, Shai Shalev-Shwartz, Omri Abend, Raz Alon, Tomer Asida, Amir Bergman, Roman Glozman, Michael Gokhman, Avshalom Manevich, Nir Ratner, Noam Rozen, Erez Shwartz, Mor Zusman, and Yoav Shoham. Jamba: A hybrid transformer-mamba language model. *arXiv preprint arXiv:2403.19887*, 2024. URL https://arxiv.org/abs/2403.19887.

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

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019.

Pascal Notin, Aaron W Kollasch, Daniel Ritter, Lood van Niekerk, Steffanie Paul, Hansen Spinner, Nathan Rollins, Ada Shaw, Ruben Weitzman, Jonathan Frazer, Mafalda Dias, Dinko Franceschi, Rose Orenbuch, Yarin Gal, and Debora S Marks. Proteingym: Large-scale benchmarks for protein design and fitness prediction. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

Zhangzhi Peng, Benjamin Schussheim, and Pranam Chatterjee. Ptm-mamba: A ptm-aware protein language model with bidirectional gated mamba blocks. *bioRxiv*, 2024.

Liliang Ren, Yang Liu, Yadong Lu, Yelong Shen, Chen Liang, and Weizhu Chen. Samba: Simple hybrid state space models for efficient unlimited context language modeling. *arXiv preprint arXiv:2406.07522*, 2024. URL https://arxiv.org/abs/2406.07522.

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

Damiano Sgarbossa, Cyril Malbranke, and Anne-Florence Bitbol. Protmamba: a homology-aware but alignment-free protein state space model. *Bioinformatics*, 41(6):btaf348, 2025. doi: 10.1093/bioinformatics/btaf348.

Jianlin Su, Yu Lu, Shengfeng Pan, Bo Wen, and Yunfeng Liu. Roformer: Enhanced transformer with rotary position embedding. *arXiv preprint arXiv:2104.09864*, 2021. URL https://arxiv.org/abs/2104.09864.

Baris E Suzek, Yuqi Wang, Hongzhan Huang, Peter B McGarvey, and Cathy H Wu. Uniref clusters: a comprehensive and scalable alternative for improving sequence similarity searches. *Bioinformatics*, 31(6):926–932, 2015. doi: 10.1093/bioinformatics/btu739.

Varun Ullanat, Bowen Jing, Samuel Sledzieski, and Bonnie Berger. Learning the language of protein-protein interactions. *Nature Communications*, 17:1199, 2026.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in Neural Information Processing Systems*, 30: 5998–6008, 2017.

Yingheng Wang, Zichen Wang, Gil Sadeh, Luca Zancato, Alessandro Achille, George Karypis, and Huzefa Rangwala. Lc-plm: Long-context protein language modeling using bidirectional mamba with shared projection layers. *arXiv preprint arXiv:2411.08909*, 2024. URL https://arxiv.org/abs/2411.08909.

Mingqian Zhu, Taiyu Hou, Ke Yuan, Yiwei Chen, Jian Chen, Yu Zhang, Ding Zhang, and Qi Liu. Hybridna: A hybrid transformer-mamba2 long-range dna language model. *arXiv preprint arXiv:2502.10807*, 2025. URL https://arxiv.org/abs/2502.10807.

Appendix

A Model details

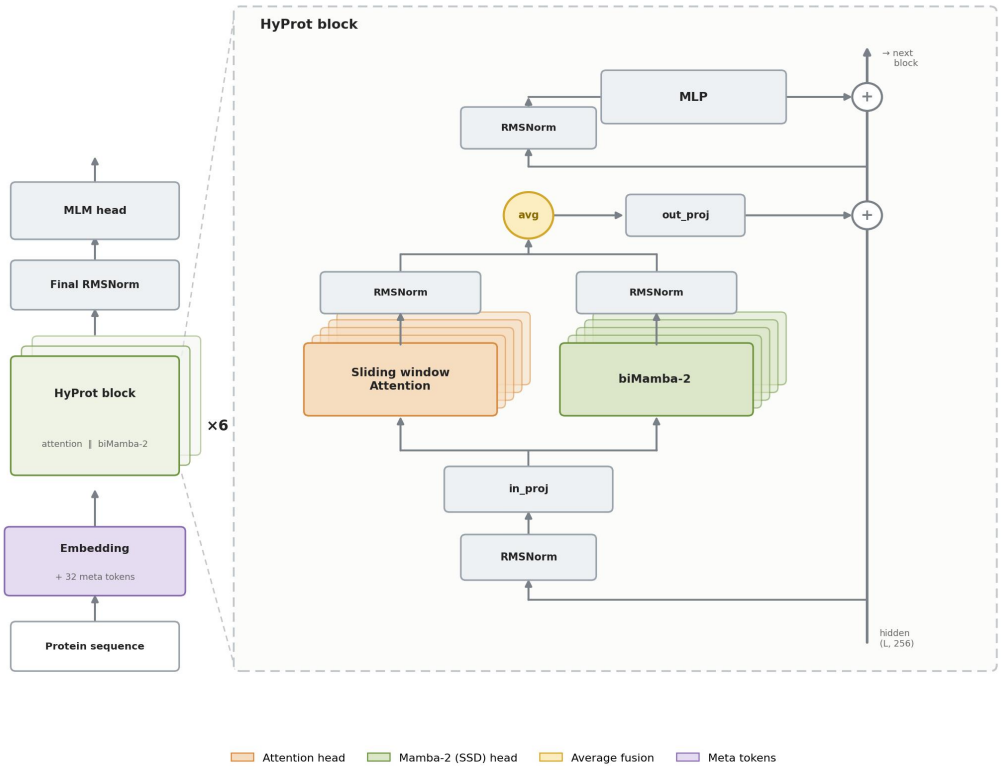


Figure 2: HyProt architecture.

Table 2: Model architecture metadata.

	HyProt	**PET**
Mixer	biMamba-2 (SSD) ‖ attention	full attention
Hidden dim	256	320
Layers	6	6
Parameters	7.62 M	7.51 M
Attention		
heads × dim	8×64	20×16
KV heads (in layer)	1	20
window	512 (1 global)	all global
SSM (Mamba-2 / SSD)		
d_{inner} (expand)	512 (2×)	–
heads × headdim	8×64	–
state size	128	–
conv kernel	3	–
FFN (act.)	384	1280
Meta tokens	32	–
Positional enc.	RoPE	RoPE

B Model Ablation

We perform ablations on components of our model, focusing on the setting of language modeling.

B.1 Meta tokens

We first start by disabling meta tokens introduced by Dong et al. (2024) to study their contribution for protein language modeling. We retrain the HyProt model from scratch with the exact same setting for the same amount of steps but without any meta token prepended, we then evaluate the ablated model on the same benchmarks as HyProt and PET (Table 3). On aggregate the two are a tie; accuracy, perplexity, and ProteinGym ρ values are all within noise.

Table 3: Pretraining metrics on validation set; ProteinGym zero-shot Spearman

	Pretraining		Zero-shot ProteinGym ρ					
Meta tokens	Acc.	PPL	Overall (217)	Binding (13)	Expr. (18)	Stab. (66)	Activity (43)	Org. fit. (77)
✓	0.2940	9.997	0.2554	0.2707	0.2790	0.3231	0.2419	0.1545
✗	0.2936	9.980	0.2537	0.2615	0.2817	0.3206	0.2421	0.1549

However, the effect is length-dependent: stratified by sequence length, the meta tokens have little impact on short sequences but lower perplexity on long inputs ($\geq$ 2k residues; Fig. 3), hence the negligible aggregate effect on ProteinGym.

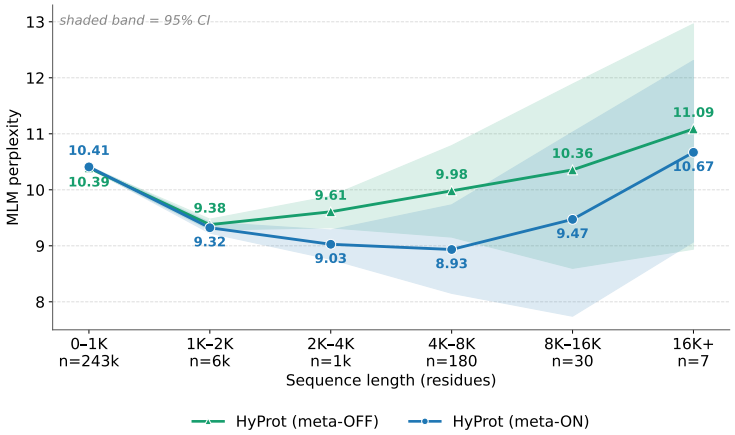


Figure 3: Comparison of MLM perplexity between HyProt and its ablation on test set by sequence length bin.

B.2 Branch Contribution

We next quantify each branch's contribution following Dong et al. (2024), by knocking out an entire branch at once (all its heads) rather than one head at a time, recording the marginal rise in MLM perplexity on the held-out set, stratified by sequence length.

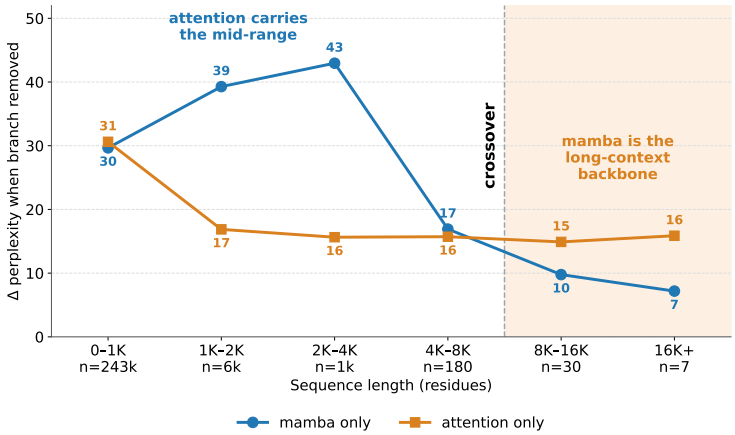


Figure 4: Variation of MLM perplexity per branch compared to HyProt on test set by sequence length bin. Lower perplexity increase is better.

The two branches specialize by length. In long contexts, the Mamba branch carries the model, as removing it raises perplexity far more than removing attention. At short-to-mid lengths the roles reverse, removing attention increases perplexity dramatically ($\sim$+42, versus $\sim$+16 for Mamba), so attention is what carries the model there. The two cross over at $\sim$8k residues. The absolute magnitudes are large since removing a whole branch is a severe perturbation; the takeaway result here is the length-dependent crossover, not the magnitudes.

C Training hyperparameters

Table 4: Pretraining hyperparameters, shared by HyProt and PET.

Optimizer	AdamW
β_1, β_2	0.9, 0.999
ϵ	10^{-8}
Weight decay	0.01
Peak learning rate	6×10^{-4}
Warmup steps	5000
Schedule	cosine to $0.1\times$ peak
Gradient clipping	1.0 (norm)
Steps	800k
Batch size	524k tokens
Total tokens	420B
Masking rate	15% (80/10/10)
Special tokens	BOS, EOS
Hardware	$2\times$ H100