HyPLM Technical Report

Hyra Team

Abstract

Protein language models (PLMs) are standard tools for variant-effect prediction, protein annotation and design, but the compute needed to train and serve them keeps growing. We present HyPLM-Flash, an efficient and open PLM trained by Hyra 2.0, the research agent developed by our team. HyPLM-Flash is a 650-million-parameter masked language model trained on 726 billion tokens, 0.73 times the data of ESM-2 650M and a fifth to an eighth of the data of E1-600M and ESM C 600M. It runs on an inference engine with fused kernels and INT8 matrix multiplication, embedding about 130,000 residues per second on a single NVIDIA H20 GPU, 1.9 to 2.7 times faster than other 600M-class encoders. In a matched evaluation against 28 public checkpoints on 32 tasks, HyPLM-Flash reaches a ProteinGym zero-shot Spearman correlation of 0.409, statistically tied with ESM-2 650M, ESM C 600M and E1-600M, and the highest organismal-fitness score among self-supervised single-sequence models. It gives the best long-range contact precision of all models evaluated, including ESM-2 15B and ESM C 6B, and a lower masked-residue loss than ESM-2 650M, ESM C 600M and E1-600M on held-out sequences and on recently released PDB chains. Matching state-of-the-art accuracy at a fraction of the training and inference cost, HyPLM-Flash sets a new Pareto frontier for protein language models. We will release the model weights and the inference engine at https://github.com/Tencent-Hunyuan/hyra-results/tree/main.

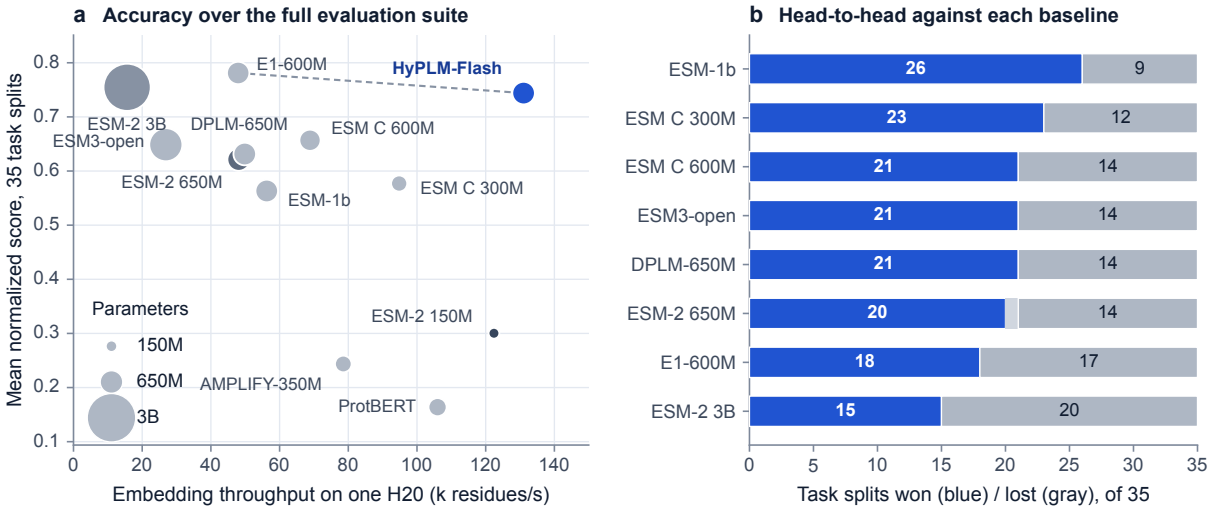


Figure 1: Accuracy and inference speed of protein language models. a, Mean score over the 35 task splits of our evaluation suite on which all plotted models have a value, against embedding throughput on one NVIDIA H20 GPU. Scores are min–max normalized on each split across the twelve models. Throughput is measured with the INT8 engine for HyPLM-Flash and with the reference implementations for the baselines. The dashed line connects Pareto-optimal models; bubble area is proportional to parameter count. **b,** Number of these splits that HyPLM-Flash wins (blue) and loses (gray) against each baseline.

1 Introduction

Protein language models (PLMs) learn the statistics of natural sequences by reconstructing masked residues, and the representations they learn capture structure, function and fitness without supervision (Rives et al., 2021; Rao et al., 2020; Lin et al., 2023). Encoders with a few hundred million parameters have become everyday tools for variant-effect prediction, protein annotation and fine-tuning, and they are increasingly run over entire proteomes and metagenomes. Their cost therefore matters. Recent encoders

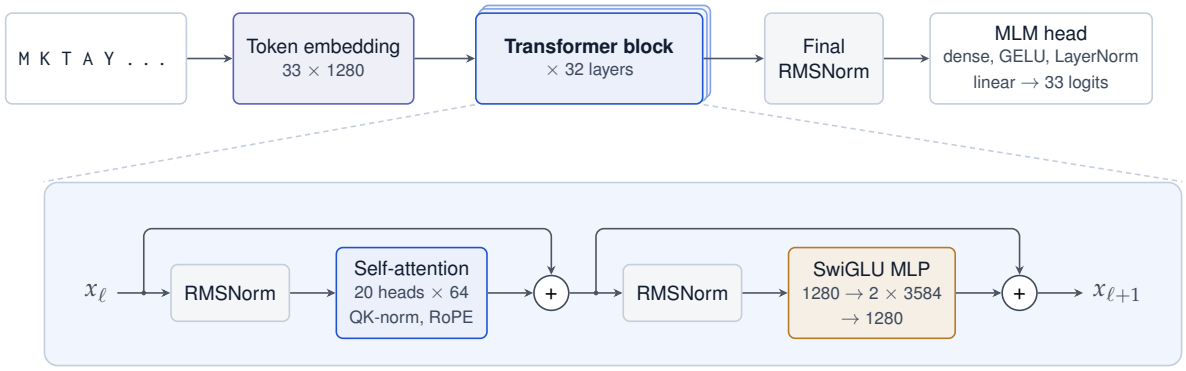


Figure 2: Architecture of HyPLM-Flash. Top: tokens from the 33-symbol ESM-2 alphabet are embedded, processed by 32 identical Transformer blocks and decoded by a masked-language-modeling head. Bottom: each block applies RMSNorm before bidirectional self-attention (per-head RMSNorm on queries and keys, rotary position embeddings) and before a SwiGLU MLP, each inside a residual connection.

of the 600M class were trained on 4 to 6 trillion tokens (ESM Team, 2024; Jain et al., 2025), four to six times the data of ESM-2, and on an NVIDIA H20 their reference implementations embed 48,000 to 69,000 residues per second, so a single pass over 2.6 billion proteins keeps an eight-GPU server busy for two to three weeks.

We present HyPLM-Flash, a protein language model trained by Hyra 2.0, the research agent developed by our team. It is as accurate as the strongest encoders of its size, and both cheaper to train and faster to serve. It keeps the vocabulary and output head of ESM-2, so it can replace ESM-2 in existing pipelines, and pairs a modern Transformer block and the Muon optimizer with a single pass over 2.6 billion curated and metagenomic sequences. A dedicated inference engine then makes full use of the GPU through padding-free batching, fused kernels and INT8 matrix multiplication.

Trained on 0.73 times the data of ESM-2 650M and a fifth to an eighth of that of E1-600M and ESM C 600M, HyPLM-Flash matches these models on zero-shot fitness prediction, leads all self-supervised single-sequence models on organismal fitness and gives the best long-range contact precision of all models we evaluated. Its engine embeds proteins 1.9 to 2.7 times faster than other encoders of the 600M class, which places HyPLM-Flash on a new Pareto frontier of accuracy and speed (Figure 1).

Our contributions are:

- HyPLM-Flash, a 650M-parameter protein language model that reaches the accuracy of the strongest encoders of its size with a fraction of their training data (Section 2).
- An inference engine for H20 GPUs that embeds about 130,000 residues per second on one GPU (Section 3).
- A matched evaluation of 28 public checkpoints from 13 model families on 32 tasks under a single protocol with paired confidence intervals (Section 4).

2 Model and Training

2.1 Architecture

HyPLM-Flash is a bidirectional Transformer encoder (Vaswani et al., 2017) trained with masked language modeling (Devlin et al., 2019). It uses the 33-token alphabet of ESM-2 and a RoBERTa-style output head (Liu et al., 2019; Lin et al., 2023), and modernizes the Transformer block with pre-normalization by RMSNorm (Zhang and Sennrich, 2019), rotary position embeddings (Su et al., 2021), per-head query–key normalization (Henry et al., 2020; Dehghani et al., 2023) and SwiGLU feed-forward layers (Shazeer, 2020) (Figure 2). Query–key normalization bounds the attention logits and keeps training stable at the high learning rates used with Muon. The model has 32 layers of width 1,280 with 20 attention heads and a SwiGLU width of 3,584, for a total of 652 million parameters, and reads proteins of up to 8,190 residues in a single window.

2.2 Training

Data. The training corpus combines UniRef90 with the UniRef50 representatives it does not contain (Suzek et al., 2015; The UniProt Consortium, 2025), MGnify90 (Richardson et al., 2023), the active entries of UniParc and the non-redundant isolate proteomes of JGI IMG (Chen et al., 2023). We keep sequences of 30 to 8,192 residues, remove exact duplicates across sources and remove, by exact match,

Table 1: Training data and compute of HyPLM-Flash and encoders of similar size.

Model	Params	Tokens	Training FLOPs 6ND	Training FLOPs Published	Corpus
Encoders of similar size					
ESM-2 650M	652M	1.0T	3.9×10^{21}	1.6×10^{22}[a]	UniRef50/90 (2021_04)
ESM-1b	652M	66B	2.6×10^{20}	—	UniRef50 (2018_03)
ESM C 600M	575M	6.2T	2.1×10^{22}	2.2×10^{22}	UniRef, MGnify, JGI (70% clusters)
E1-600M	641M	4.0T	1.5×10^{22}	—	Profluent Protein Atlas, UniRef
AMPLIFY-350M	354M	2.5T[b]	5.3×10^{21}	4.6×10^{21}	UniRef100, OAS, SCOP (UR100P)
DPLM-650M	652M	100B	4.3×10^{21}[c]	—	UniRef50
ESM3-open	1.4B	300B	2.5×10^{21}	—	UniRef, MGnify, JGI, structures
HyPLM-Flash	652M	726B	2.8×10^{21}	—	UniRef90/50, MGnify90, UniParc, JGI; 1.6M structures

6ND: total parameters N times training tokens D; published FLOPs as reported by the authors; — not published. [a] Estimated by the authors from GPU-days, not a 6ND count. [b] Upper bound: counts padded sequence slots. [c] DPLM-650M continues ESM-2 650M; tokens cover the final stage and FLOPs include the parent model (final stage alone: 3.9×10^{20}).

Sources: ESM-2 (Lin et al., 2023; Akin, 2022), ESM-1b (Rives et al., 2021; Meta AI, 2022), ESM C (Candido et al., 2026; ESM Team, 2024), E1 (Jain et al., 2025), AMPLIFY (Fournier et al., 2024), DPLM (Wang et al., 2024), ESM3 (Hayes et al., 2025).

the sequences of the benchmarks used during development, including all ProteinGym wild types and variants. The resulting 2.6 billion sequences (659 billion residues) are sampled in proportion to their length, so that metagenomic and archival sequences supply about 80% of the training tokens, and each source is seen about once.

Objective and optimization. We use the standard masking recipe: 15% of the residues are selected, of which 80% are replaced by a mask token, 10% by a random amino acid and 10% are left unchanged. The weight matrices of the Transformer blocks are trained with Muon (Jordan et al., 2024), using decoupled weight decay and an update scale matched to AdamW (Liu et al., 2025), and all other parameters with AdamW (Loshchilov and Hutter, 2019). The learning rate follows a warmup–stable–decay schedule (Hu et al., 2024; Wen et al., 2024).

Training stages. Pretraining covers 713 billion tokens on crops of up to 1,022 residues. A second stage of 13 billion tokens extends the crops to 8,190 residues and adds auxiliary structure objectives on 1.6 million protein chains with experimental (Berman et al., 2000) or predicted (Lin et al., 2023; Varadi et al., 2022) structures: a residue head predicts secondary structure (Kabsch and Sander, 1983) and Foldseek 3Di states (van Kempen et al., 2024), and a pair head predicts Cβ distances (Jumper et al., 2021). The auxiliary heads are used only during training. Before training, we removed every structure chain with an MMseqs2 hit (E-value $\leq 10^{-3}$, sequence identity $\geq 30\%$) to the test proteins of the structure and likelihood tasks or to a ProteinGym wild type.

Training compute. In total, HyPLM-Flash was trained on 726 billion tokens, or 2.8×10^{21} FLOPs by the usual $6ND$ estimate (Kaplan et al., 2020; Hoffmann et al., 2022). This is 0.73 times the data of ESM-2 650M and 5.5 and 8.5 times less than E1-600M and ESM C 600M, whose 4.0 and 6.2 trillion tokens amount to 1.5×10^{22} and 2.1×10^{22} FLOPs (Table 1).

3 Inference Engine

Embedding proteomes and scoring variant libraries are dominated by inference, and reference implementations of protein language models leave much of a modern GPU idle. We serve HyPLM-Flash with an inference engine written for NVIDIA H20 GPUs, which offer 148 TFLOPS in BF16 and 296 TOPS in INT8 at 4.0 TB/s of memory bandwidth.

The engine rests on three ideas. Proteins are packed into a single token stream without padding, which would otherwise fill 42% of the slots of a typical batch, and attention runs in the variable-length mode of FlashAttention-3 (Shah et al., 2024). Three Triton kernels (Tillet et al., 2019) absorb every memory-bound operation: the residual addition with RMSNorm, query–key normalization with the rotary embedding, and the SwiGLU activation, each fused with the quantization of its output, while the residual stream

stays in FP32. Finally, the four projections of every block run as INT8 matrix multiplications. Activations are scaled per token and per group of 128 channels, which confines the heavy-tailed outliers of the feed-forward activations to their own group, and weights are scaled per output channel; each 128-wide slice of the reduction accumulates in INT32 and is promoted to FP32 with its activation scale, the strategy DeepSeek-V3 uses for FP8 (DeepSeek-AI, 2024).

On a common workload of 2,048 held-out proteins, the engine embeds 131,000 residues per second on one H20. This is 1.9 to 2.7 times faster than ESM C 600M, ESM-1b, ESM-2 650M, DPLM-650M and E1-600M in their reference implementations, and faster than ESM-2 150M, which has a quarter of the parameters (Figure 1a). Accuracy in this report is evaluated with BF16 arithmetic. INT8 keeps zero-shot fitness, structure, function annotation and likelihood within 0.01 of these results; for supervised regression on near-identical protein variants, whose representations differ by only about 1%, we recommend the BF16 mode of the engine, which runs at 90,000 residues per second.

4 Evaluation

We compare HyPLM-Flash with 28 public checkpoints from 13 model families, from ESM-2 8M to ESM-2 15B (29 models in all), on 32 tasks. All models run through the same evaluation code in BF16. Zero-shot scores come from the masked-language-model head, and every supervised task fits a linear probe on frozen representations from the layer, among six relative depths, that performs best on validation data. Confidence intervals are paired bootstrap intervals over assays, proteins or sequences. The main text focuses on encoders of similar size; Appendix A describes the protocols and Appendix B reports every task for HyPLM-Flash and twelve other encoders. E1 (Jain et al., 2025) is retrieval-augmented: it is trained with homologous sequences in its context and improves when they are supplied, at the cost of an external sequence database, a homology search for every query and longer inputs. As for every other model, we run E1-600M on the query sequence alone.

4.1 Zero-shot fitness prediction

We score each substitution by the masked-marginal log-likelihood ratio (Meier et al., 2021) and report the Spearman correlation over the 217 substitution assays of ProteinGym v1.3 (Notin et al., 2023) with the benchmark's official aggregation. Our runs of eight published ESM-2 and ESM C checkpoints reproduce the leaderboard to within 0.0025.

HyPLM-Flash reaches a Spearman correlation of 0.409 (Figure 3a). Its differences to ESM-2 650M (-0.006, 95% confidence interval $[-0.025, 0.010]$), ESM C 600M ($+0.003$, $[-0.015, 0.019]$), DPLM-650M (-0.001, $[-0.019, 0.015]$) and E1-600M (-0.011, $[-0.029, 0.005]$) are not significant, and neither are those to ESM-2 3B, ESM-2 15B and ESM C 6B, which were trained with 6 to 83 times more compute. Among the 55 single-sequence entries of the public leaderboard, HyPLM-Flash is tied for eighth place (Table 5); the two entries at the top, VespaG and VESPA, are supervised predictors built on PLM embeddings (Marquet et al., 2022; 2024). By selection type (Figure 3b), HyPLM-Flash scores 0.406 on organismal fitness, the highest of all self-supervised single-sequence models on the leaderboard, and is on par with the other encoders on activity and expression.

4.2 Frozen-representation tasks

Following TAPE (Rao et al., 2019) and Rao et al. (2020), we probe frozen representations with a linear per-residue probe for secondary structure, a 32-dimensional bilinear head on residue pairs for contacts, and linear probes on mean-pooled representations for all other tasks (Table 2).

Structure. HyPLM-Flash gives the best long-range contact precision of all 29 models: a precision at $L/5$ of 0.650 on ProteinNet, against 0.392 for ESM-2 650M, 0.434 for ESM C 600M and 0.645 for E1-600M, and the highest scores on the CASP14 and CASP15 targets as well (Appendix B), ahead of ESM-2 15B and ESM C 6B. Its three-state secondary-structure accuracy on CB513 (0.882) is above that of ESM-2 650M (0.844) and ESM C 600M (0.851), on par with E1-600M (0.880), and exceeded only by ESM3-open (0.896) and ESM C 6B (0.884) among all models.

Function, localization and fitness. On enzyme and Gene Ontology annotation, HyPLM-Flash is within one point of the best encoder of its size (EC $F_{\max}$ 0.885, GO-MF 0.661). On fitness and property regression it ranks first among the models of Table 2 on AAV (two-vs-many, $\rho = 0.650$) and amylase (0.696), and second on Meltome (0.672) and fluorescence (0.669). Averaged over the ten regression splits used for summary statistics (Appendix A), it scores 0.606, the highest of these models, ahead of E1-600M (0.597), ESM C 600M (0.587), ESM-2 650M (0.572) and the 4.4 times larger ESM-2 3B (0.573).

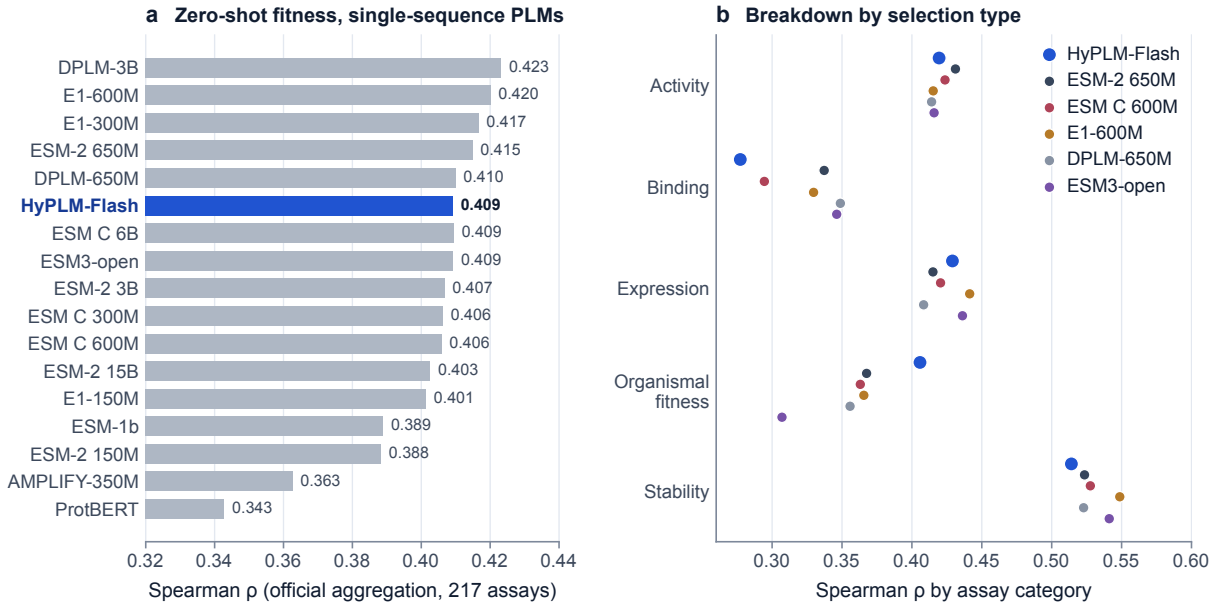


Figure 3: ProteinGym zero-shot fitness prediction. a, Average Spearman correlation over 217 substitution assays (official aggregation) for single-sequence models; ESM3-open is scored from sequence alone. **b,** Correlation by selection type for HyPLM-Flash and the main encoders of similar size.

Table 2: Frozen-representation tasks. Test metrics at the validation-selected layer: three-state accuracy on CB513 (SS), long-range contact precision at $L/5$ on ProteinNet, fold accuracy, F_{max} for EC and GO molecular function, 10-class localization accuracy, and Spearman ρ for AAV (two-vs-many), amylase (one-to-many), fluorescence, Meltome (mixed split) and stability. Bold marks the best value in each column.

	Structure			Function			Fitness and properties				
Model	SS (Q3)	Contact	Fold	EC	GO-MF	Loc.	AAV	Amylase	Fluor.	Meltome	Stability
ESM-2 650M	0.844	0.392	0.373	0.888	**0.667**	0.806	0.518	0.640	0.664	0.633	0.639
ESM C 600M	0.851	0.434	0.345	0.886	0.665	0.808	0.551	0.681	0.663	0.670	0.706
E1-600M	0.880	0.645	0.373	0.881	0.661	**0.839**	0.583	0.579	0.660	**0.690**	0.712
DPLM-650M	0.842	0.394	0.370	0.882	0.663	0.806	0.569	0.674	0.668	0.623	0.746
ESM-1b	0.814	0.410	0.362	**0.889**	0.659	0.793	0.544	0.653	0.659	0.593	0.596
AMPLIFY-350M	0.797	0.210	0.286	0.865	0.631	0.729	0.574	0.592	0.660	0.585	0.579
ESM3-open	**0.896**	0.568	**0.408**	0.880	0.664	0.750	0.551	0.618	**0.669**	0.573	**0.758**
HyPLM-Flash	0.882	**0.650**	0.322	0.885	0.661	0.794	**0.650**	**0.696**	0.669	0.672	0.687

4.3 Masked-residue likelihood

We compute the negative log-likelihood (NLL) of masked residues with identical masks for all models on two panels: 479 proteins from our held-out UniRef split, and 887 PDB chains released after 2 September 2026, later than the training data of every model evaluated, clustered at 50% identity with MMseqs2 (Steinegger and Söding, 2017) after removing exact matches to our training corpus. HyPLM-Flash has a lower NLL than ESM-2 650M, ESM C 600M and E1-600M on both panels (Table 3). On the recent PDB chains it is also ahead of ESM-2 3B and ESM-2 15B (1.148 against 1.327 and 1.162); of all models for which we report a likelihood, only ESM C 6B and ProtBERT score lower.

5 Conclusion

HyPLM-Flash shows that a protein language model can reach the accuracy of the strongest encoders of its size with a fraction of their training data, and that an inference stack built around the target GPU makes it the fastest of them. It matches ESM-2 650M, ESM C 600M and E1-600M on zero-shot fitness prediction, leads all self-supervised single-sequence models on organismal fitness, gives the best contact precision in our evaluation and models held-out and recent sequences better than ESM-2 650M, ESM C 600M and E1-600M, while embedding proteins 1.9 to 2.7 times faster than these models. Together, these results place HyPLM-Flash on a new Pareto frontier of accuracy and efficiency.

Table 3: Masked-residue likelihood. Negative log-likelihood (NLL, nats) and perplexity (PPL) on held-out UniRef proteins and on PDB chains released after 2 September 2026. Lower is better.

	Held-out UniRef proteins		Recent PDB chains	
Model	NLL	PPL	NLL	PPL
Encoders of similar size				
ESM-2 650M	1.820	6.17	1.591	4.91
ESM C 600M	1.720	5.59	1.478	4.39
E1-600M	1.811	6.11	1.560	4.76
AMPLIFY-350M	1.743	5.72	1.155	3.17
DPLM-650M	1.961	7.10	1.796	6.03
ProtBERT	1.754	5.77	1.105	3.02
Larger encoders				
ESM-2 3B	1.667	5.30	1.327	3.77
ESM-2 15B	1.561	4.76	1.162	3.19
ESM C 6B	**1.414**	**4.11**	**1.054**	**2.87**
HyPLM-Flash	1.698	5.46	1.148	3.15

Outlook. HyPLM-Flash is the first release of the HyPLM family. Hyra 2.0 continues to develop it, and we will release updated versions as data and training recipes improve. We are also training HyPLM-Pro, a larger model aimed at the highest accuracy rather than speed. Like other single-sequence models, HyPLM-Flash reads sequence alone; conditioning on structure or on retrieved homologs (Su et al., 2024; Jain et al., 2025) is a natural direction for future versions.

Author Information

Haowei Lin, Cai Zhou and Xiangxin Zhou. All authors are affiliated with Tencent Hunyuan.

References

Halil Akin. Pretraining resources of ESM2. GitHub discussion #414, facebookresearch/esm, December 6, 2022. URL https://github.com/facebookresearch/esm/discussions/414.

Helen M. Berman, John Westbrook, Zukang Feng, Gary Gilliland, T. N. Bhat, Helge Weissig, Ilya N. Shindyalov, and Philip E. Bourne. The protein data bank. *Nucleic Acids Research*, 28(1):235–242, 2000. doi: 10.1093/nar/28.1.235.

Salvatore Candido, Thomas Hayes, Alexander Derry, Roshan Rao, Zeming Lin, Robert Verkuil, Bryan Z. Wu, Jin Sub Lee, Elise S. Bruguera, Jehan A. Keval, Mykhailo Kopylov, John E. Pak, Wesley Wu, Neil Thomas, Samson Mataraso, Alvin Hsu, Ashton C. Trotman-Grant, Kilian Fatras, Allan dos Santos Costa, Rohil Badkundri, Halil Akın, Deniz Oktay, Jonathan Deaton, Elizabeth Montabana, Hrishita Sitwala, Yue Yu, Marius Wiggert, Dylan Alexander Carlin, Anthony W. Goering, Tomasz Blazejewski, McCullen Sandora, Michael Hla, Tina Z. Jia, Leon H. Kloker, Nicholas J. Sofroniew, Masatoshi Uehara, Jassi Pannu, Sharrol Bachas, Daniel S. Liu, Tom Sercu, and Alexander Rives. Language modeling materializes a world model of protein biology. *bioRxiv*, 2026. doi: 10.64898/2026.06.03.729735.

I-Min A. Chen, Ken Chu, Krishnaveni Palaniappan, Anna Ratner, Jinghua Huang, Marcel Huntemann, Patrick Hajek, Stephan J. Ritter, Cody Webb, Dongying Wu, Neha J. Varghese, T. B. K. Reddy, Supratim Mukherjee, Galina Ovchinnikova, Matt Nolan, Rekha Seshadri, Simon Roux, Axel Visel, Tanja Woyke, Emiley A. Eloe-Fadrosh, Nikos C. Kyrpides, and Natalia N. Ivanova. The IMG/M data management and analysis system v.7: content updates and new features. *Nucleic Acids Research*, 51(D1):D723–D732, 2023. doi: 10.1093/nar/gkac976.

Christian Dallago, Jody Mou, Kadina E. Johnston, Bruce J. Wittmann, Nicholas Bhattacharya, Samuel Goldman, Ali Madani, and Kevin K. Yang. FLIP: Benchmark tasks in fitness landscape inference for proteins. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, 2021.

DeepSeek-AI. DeepSeek-V3 technical report. *arXiv preprint arXiv:2412.19437*, 2024.

Mostafa Dehghani, Josip Djolonga, Basil Mustafa, Piotr Padlewski, Jonathan Heek, Justin Gilmer, Andreas Steiner, Mathilde Caron, Robert Geirhos, Ibrahim Alabdulmohsin, Rodolphe Jenatton, Lucas Beyer, Michael Tschannen, Anurag Arnab, Xiao Wang, Carlos Riquelme, Matthias Minderer, Joan Puigcerver,

Utku Evci, Manoj Kumar, Sjoerd van Steenkiste, Gamaleldin F. Elsayed, Aravindh Mahendran, Fisher Yu, Avital Oliver, Fantine Huot, Jasmijn Bastings, Mark Patrick Collier, Alexey Gritsenko, Vighnesh Birodkar, Cristina Vasconcelos, Yi Tay, Thomas Mensink, Alexander Kolesnikov, Filip Pavetić, Dustin Tran, Thomas Kipf, Mario Lučić, Xiaohua Zhai, Daniel Keysers, Jeremiah Harmsen, and Neil Houlsby. Scaling vision transformers to 22 billion parameters. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 7480–7512, 2023.

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

ESM Team. ESM Cambrian: Revealing the mysteries of proteins with unsupervised learning. EvolutionaryScale website, December 4, 2024. URL https://www.evolutionaryscale.ai/blog/esm-cambrian.

Quentin Fournier, Robert M. Vernon, Almer van der Sloot, Benjamin Schulz, Sarath Chandar, and Christopher James Langmead. Protein language models: Is scaling necessary? *bioRxiv*, 2024. doi: 10.1101/2024.09.23.614603.

Thomas Hayes, Roshan Rao, Halil Akin, Nicholas J. Sofroniew, Deniz Oktay, Zeming Lin, Robert Verkuil, Vincent Q. Tran, Jonathan Deaton, Marius Wiggert, Rohil Badkundri, Irhum Shafkat, Jun Gong, Alexander Derry, Raul S. Molina, Neil Thomas, Yousuf A. Khan, Chetan Mishra, Carolyn Kim, Liam J. Bartie, Matthew Nemeth, Patrick D. Hsu, Tom Sercu, Salvatore Candido, and Alexander Rives. Simulating 500 million years of evolution with a language model. *Science*, 387(6736):850–858, 2025. doi: 10.1126/science.ads0018.

Alex Henry, Prudhvi Raj Dachapally, Shubham Shantaram Pawar, and Yuxuan Chen. Query-key normalization for transformers. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4246–4253, 2020. doi: 10.18653/v1/2020.findings-emnlp.379.

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, Tom Hennigan, Eric Noland, Katie Millican, George van den Driessche, Bogdan Damoc, Aurelia Guy, Simon Osindero, Karen Simonyan, Erich Elsen, Jack W. Rae, Oriol Vinyals, and Laurent Sifre. Training compute-optimal large language models. In *Advances in Neural Information Processing Systems*, volume 35, 2022.

Shengding Hu, Yuge Tu, Xu Han, Chaoqun He, Ganqu Cui, Xiang Long, Zhi Zheng, Yewei Fang, Yuxiang Huang, Weilin Zhao, Xinrong Zhang, Zheng Leng Thai, Kaihuo Zhang, Chongyi Wang, Yuan Yao, Chenyang Zhao, Jie Zhou, Jie Cai, Zhongwu Zhai, Ning Ding, Chao Jia, Guoyang Zeng, Dahai Li, Zhiyuan Liu, and Maosong Sun. MiniCPM: Unveiling the potential of small language models with scalable training strategies. *arXiv preprint arXiv:2404.06395*, 2024.

Sarthak Jain, Joel Beazer, Jeffrey A. Ruffolo, Aadyot Bhatnagar, and Ali Madani. E1: Retrieval-augmented protein encoder models. *bioRxiv*, 2025. doi: 10.1101/2025.11.12.688125.

Keller Jordan, Yuchen Jin, Vlado Boza, Jiacheng You, Franz Cesista, Laker Newhouse, and Jeremy Bernstein. Muon: An optimizer for hidden layers in neural networks. Blog post, 2024. URL https://kellerjordan.github.io/posts/muon/.

John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, et al. Highly accurate protein structure prediction with AlphaFold. *Nature*, 596:583–589, 2021. doi: 10.1038/s41586-021-03819-2.

Wolfgang Kabsch and Christian Sander. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers*, 22(12):2577–2637, 1983. doi: 10.1002/bip.360221211.

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, 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.

Jingyuan Liu, Jianlin Su, Xingcheng Yao, Zhejun Jiang, Guokun Lai, Yulun Du, Yidao Qin, Weixin Xu, Enzhe Lu, Junjie Yan, Yanru Chen, Huabin Zheng, Yibo Liu, Shaowei Liu, Bohong Yin, Weiran He, Han Zhu, Yuzhi Wang, Jianzhou Wang, Mengnan Dong, Zheng Zhang, Yongsheng Kang, Hao Zhang, Xinran Xu, Yutao Zhang, Yuxin Wu, Xinyu Zhou, and Zhilin Yang. Muon is scalable for LLM training. *arXiv preprint arXiv:2502.16982*, 2025.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. RoBERTa: A robustly optimized BERT pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.

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

Céline Marquet, Michael Heinzinger, Tobias Olenyi, Christian Dallago, Kyra Erckert, Michael Bernhofer, Dmitrii Nechaev, and Burkhard Rost. Embeddings from protein language models predict conservation and variant effects. *Human Genetics*, 141:1629–1647, 2022. doi: 10.1007/s00439-021-02411-y.

Céline Marquet, Julius Schlensok, Marina Abakarova, Burkhard Rost, and Elodie Laine. Expert-guided protein language models enable accurate and blazingly fast fitness prediction. *Bioinformatics*, 40(11): btae621, 2024. doi: 10.1093/bioinformatics/btae621.

Joshua Meier, Roshan Rao, Robert Verkuil, Jason Liu, Tom Sercu, and Alexander Rives. Language models enable zero-shot prediction of the effects of mutations on protein function. *bioRxiv*, 2021. doi: 10.1101/2021.07.09.450648. Published at NeurIPS 2021.

Meta AI. facebook/esm1b_t33_650M_UR50S model card. Hugging Face, 2022. URL https://huggingface.co/facebook/esm1b_t33_650M_UR50S.

Pascal Notin, Aaron W. Kollasch, Daniel Ritter, Lood van Niekerk, Steffanie Paul, Hansen Spinner, Nathan Rollins, Ada Shaw, Rose Orenbuch, Ruben Weitzman, Jonathan Frazer, Mafalda Dias, Dinko Franceschi, Yarin Gal, and Debora S. Marks. ProteinGym: Large-scale benchmarks for protein fitness prediction and design. In *Advances in Neural Information Processing Systems*, volume 36, pages 64331–64379, 2023.

Roshan Rao, Nicholas Bhattacharya, Neil Thomas, Yan Duan, Xi Chen, John Canny, Pieter Abbeel, and Yun S. Song. Evaluating protein transfer learning with TAPE. In *Advances in Neural Information Processing Systems*, volume 32, pages 9689–9701, 2019.

Roshan Rao, Joshua Meier, Tom Sercu, Sergey Ovchinnikov, and Alexander Rives. Transformer protein language models are unsupervised structure learners. *bioRxiv*, 2020. doi: 10.1101/2020.12.15.422761. Published at ICLR 2021.

Lorna Richardson, Ben Allen, Germana Baldi, Martin Beracochea, Maxwell L. Bileschi, Tony Burdett, Josephine Burgin, Juan Caballero-Pérez, Guy Cochrane, Lucy J. Colwell, Tom Curtis, Alejandra Escobar-Zepeda, Tatiana A. Gurbich, Varsha Kale, Anton Korobeynikov, Shriya Raj, Alexander B. Rogers, Ekaterina Sakharova, Santiago Sanchez, Darren J. Wilkinson, and Robert D. Finn. MGnify: the microbiome sequence data analysis resource in 2023. *Nucleic Acids Research*, 51(D1):D753–D759, 2023. doi: 10.1093/nar/gkac1080.

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.

Jay Shah, Ganesh Bikshandi, Ying Zhang, Vijay Thakkar, Pradeep Ramani, and Tri Dao. FlashAttention-3: Fast and accurate attention with asynchrony and low-precision. In *Advances in Neural Information Processing Systems*, volume 37, pages 68658–68685, 2024.

Noam Shazeer. GLU variants improve transformer. *arXiv preprint arXiv:2002.05202*, 2020.

Martin Steinegger and Johannes Söding. MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nature Biotechnology*, 35:1026–1028, 2017. doi: 10.1038/nbt.3988.

Jianlin Su, Yu Lu, Shengfeng Pan, Ahmed Murtadha, Bo Wen, and Yunfeng Liu. RoFormer: Enhanced transformer with rotary position embedding. *arXiv preprint arXiv:2104.09864*, 2021.

Jin Su, Chenchen Han, Yuyang Zhou, Junjie Shan, Xibin Zhou, and Fajie Yuan. SaProt: Protein language modeling with structure-aware vocabulary. In *International Conference on Learning Representations (ICLR)*, 2024.

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

The UniProt Consortium. UniProt: the universal protein knowledgebase in 2025. *Nucleic Acids Research*, 53(D1):D609–D617, 2025. doi: 10.1093/nar/gkae1010.

Philippe Tillet, H. T. Kung, and David Cox. Triton: An intermediate language and compiler for tiled neural network computations. In *Proceedings of the 3rd ACM SIGPLAN International Workshop on Machine Learning and Programming Languages (MAPL)*, pages 10–19, 2019. doi: 10.1145/3315508.3329973.

Michel van Kempen, Stephanie S. Kim, Charlotte Tumescheit, Milot Mirdita, Jeongjae Lee, Cameron L. M. Gilchrist, Johannes Söding, and Martin Steinegger. Fast and accurate protein structure search with Foldseek. *Nature Biotechnology*, 42:243–246, 2024. doi: 10.1038/s41587-023-01773-0.

Mihaly Varadi, Stephen Anyango, Mandar Deshpande, Sreenath Nair, Cindy Natassia, Galabina Yordanova, David Yuan, Oana Stroe, Gemma Wood, Agata Laydon, et al. AlphaFold protein structure database: massively expanding the structural coverage of protein-sequence space with high-accuracy models. *Nucleic Acids Research*, 50(D1):D439–D444, 2022. doi: 10.1093/nar/gkab1061.

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

Xinyou Wang, Zaixiang Zheng, Fei Ye, Dongyu Xue, Shujian Huang, and Quanquan Gu. Diffusion language models are versatile protein learners. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 52309–52333, 2024.

Kaiyue Wen, Zhiyuan Li, Jason Wang, David Hall, Percy Liang, and Tengyu Ma. Understanding warmup-stable-decay learning rates: A river valley loss landscape perspective. *arXiv preprint arXiv:2410.05192*, 2024.

Jacob West-Roberts, Joshua Kravitz, Nishant Jha, Andre Cornman, and Yunha Hwang. Diverse genomic embedding benchmark for functional evaluation across the tree of life. *bioRxiv*, 2024. doi: 10.1101/2024.07.10.602933.

Minghao Xu, Zuobai Zhang, Jiarui Lu, Zhaocheng Zhu, Yangtian Zhang, Chang Ma, Runcheng Liu, and Jian Tang. PEER: A comprehensive and multi-task benchmark for protein sequence understanding. In *Advances in Neural Information Processing Systems*, volume 35, pages 35156–35173, 2022.

Biao Zhang and Rico Sennrich. Root mean square layer normalization. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

A Evaluation Protocols

Models. Baselines are public checkpoints run in BF16 through thin adapters around their reference implementations: Hugging Face Transformers for ESM-1b, ESM-2, DPLM, SaProt, ProtBERT, ProtT5 and Ankh; the `esm` package for ESM C and ESM3; and the authors' code for E1 (single-sequence mode, without retrieved context sequences), AMPLIFY and xTrimoPGLM. Inputs longer than a model's window are processed in windows with 50% overlap whose residue representations are averaged.

Throughput. For the speed comparison (Figure 1a), all models embed the same 2,048 held-out proteins (639,806 residues) in length-sorted batches of at most 16,384 tokens on one H20, using the fastest attention kernel each implementation supports: FlashAttention-2 for ESM-1b, ESM-2 and DPLM, FlashAttention for ESM C and E1, xformers for AMPLIFY and PyTorch SDPA for ProtBERT and ESM3-open. HyPLM-Flash runs on its INT8 engine.

Probes. Sequence-level tasks use representations mean-pooled over residues without special tokens; per-residue tasks use the residue representations. Features are cached in FP16, or in FP32 for the Ankh models, whose activations exceed the FP16 range. For every task and model we evaluate six layers at relative depths 0.25, 0.5, 0.625, 0.75, 0.875 and 1.0 and select the layer and the regularization strength on the validation split, or by 5-fold cross-validation on the training split where a task has none (DGEB EC). Regression probes are closed-form ridge regressions and classification probes are L2-regularized logistic regressions fitted with L-BFGS; multi-label tasks fit one logistic output per term and report $F_{\max}$, and the HumanPPI pair task uses the features $[a, b, a \odot b, |a - b|]$. The contact head projects residue features to 32 dimensions and scores pairs with a symmetric bilinear form; it is trained with AdamW on up to 8,000 ProteinNet proteins. Contacts are residue pairs whose Cβ atoms (Cα for ProteinNet) are closer than 8 Å, and long range means a sequence separation of at least 24. DGEB Arch retrieval is zero-shot, by cosine similarity between mean-pooled representations.

Zero-shot scores and likelihood. A variant's ProteinGym score is the sum over its substitutions of $\log p(\text{mutant}) - \log p(\text{wild type})$, with each position masked on its own in the wild-type sequence and, for long targets, one window centered on the position. Per-assay Spearman correlations are rounded to three decimals and aggregated with the official procedure. The likelihood panels use fixed masks (15% of the canonical residues, one seeded draw per protein) and identical crops for every model. Models without a suitable masked-language-model head (ProtT5, Ankh, Ankh3) are omitted from both.

Statistics. Confidence intervals use 5,000 bootstrap replicates over assays (recomputing the official aggregation for each replicate), proteins (structure and likelihood), queries (retrieval), or sequences or pairs; paired differences use identical resamples. Three regression splits, GB1 two-vs-rest and three-vs-rest and the cross-protein hydrolase split of FLIP2, are reported in Tables 6 and 7 but left out of summary statistics (Figure 1 and the regression average of Section 4.2): their validation data come from the training distribution while the test sets are strongly shifted, so probe selection on them is dominated by noise, and numerically equivalent runs of the same model differ by up to 0.2.

Deviations. We could not reproduce the published ProteinGym scores of xTrimoPGLM: our runs fall 0.07 to 0.09 below the leaderboard, and its likelihood is implausibly high, which points to an incompatibility between our adapter and its masked-language-model head; we therefore withhold its zero-shot and likelihood results, while its embedding-based results are unaffected. ESM3-open and SaProt are evaluated from sequence alone, whereas their published ProteinGym results also use structure inputs.

B Additional Results

Table 5 lists the top of the ProteinGym leaderboard, and Tables 6 and 7 the primary metric of every task for HyPLM-Flash and twelve other encoders.

Table 4: Evaluation tasks. Split sizes in proteins (ProteinGym: assays; CASP: domains).

Task	Metric	Train / valid / test	Source
Zero-shot			
ProteinGym substitutions	Spearman ρ	217 assays	ProteinGym v1.3 (Notin et al., 2023)
Held-out panel	NLL	– / – / 479	held-out UniRef split
PDB time split	NLL	– / – / 887	RCSB PDB entries released 9 Sep to 7 Oct 2026
Structure			
Secondary structure (Q3)	accuracy	8,678 / 2,170 / 513	TAPE (Rao et al., 2019); CB513, TS115, CASP12
Contacts, ProteinNet	P@L/5 long	25,299 / 224 / 40	TAPE, ProteinNet CASP12
Contacts, CASP14 / CASP15	P@L/5 long	– / – / 82, 95	CASP evaluation domains
Remote homology	fold accuracy	12,312 / 736 / 718	TAPE (SCOP)
Function, localization and interaction			
EC (two splits)	$F_{\max}$	13,090 / 1,465 / 1,604	SaProt EC split; DeepFRI EC split
GO-MF, GO-BP, GO-CC	$F_{\max}$	26,225 / 2,904 / 3,350	DeepFRI GO split as packaged by SaProt
DGEB EC	macro F1	512 / – / 128	DGEB (West-Roberts et al., 2024)
DGEB Arch retrieval	MAP@5	2,343 queries	DGEB
DeepLoc (10-class, binary)	accuracy	8,743 / 2,190 / 2,745	DeepLoc 1.0 via PEER (Xu et al., 2022)
Metal-ion binding	accuracy	5,068 / 662 / 665	as packaged by SaProt
HumanPPI	accuracy	26,319 / 234 / 180	PEER
Fitness and property regression			
AAV (two-vs-many, low-vs-high)	Spearman ρ	28,626 / 3,181 / 50,776	FLIP (Dallago et al., 2021)
GB1 (three splits)	Spearman ρ	381 / 43 / 8,309	FLIP
Meltome (human cell, mixed)	Spearman ρ	5,149 / 643 / 1,366	FLIP
Amylase (one-to-many, by mutation count)	Spearman ρ	2,574 / 644 / 488	FLIP2
Hydrolase to P06241	Spearman ρ	11,970 / 2,993 / 9,972	FLIP2 (cross-protein)
Fluorescence, stability	Spearman ρ	21,446 / 5,362 / 27,217	TAPE
Thermostability	Spearman ρ	5,056 / 639 / 1,336	Meltome human cell as packaged by SaProt

Where a row groups several tasks, sizes are for the first.

Table 5: Top of the ProteinGym zero-shot leaderboard (substitutions, single-sequence models, Spearman ρ by selection type). Models are ranked by their average at the precision shown, and tied models share a rank. † our evaluation; other rows are published leaderboard values (ProteinGym repository, commit 144fe22). VespaG and VESPA are supervised (Marquet et al., 2022; 2024).

Rank	Model	Activity	Binding	Expression	Organismal fitness	Stability	Average
1	VespaG	**0.493**	**0.370**	**0.456**	0.437	0.533	**0.458**
2	VESPA	0.468	0.366	0.404	**0.439**	0.500	0.435
3	DPLM-3B†	0.427	0.346	0.422	0.398	0.523	0.423
4	E1-600M†	0.415	0.330	0.441	0.366	**0.549**	0.420
5	E1-300M†	0.438	0.332	0.429	0.346	0.537	0.417
6	ESM2 (650M)	0.425	0.337	0.415	0.368	0.523	0.414
7	DPLM-650M†	0.414	0.349	0.408	0.356	0.523	0.410
8	**HyPLM-Flash**†	0.419	0.277	0.429	0.406	0.514	0.409
8	ESM C 6B†	0.428	0.309	0.417	0.374	0.518	0.409
10	ESM-1v (ensemble)	0.420	0.320	0.429	0.386	0.477	0.406
10	ESM-C (300M)	0.423	0.315	0.408	0.360	0.526	0.406
10	ESM2 (3B)	0.417	0.321	0.403	0.378	0.509	0.406
13	ESM-C (600M)	0.423	0.294	0.420	0.362	0.528	0.405
14	E1-150M†	0.426	0.325	0.420	0.304	0.532	0.401

Table 6: All tasks, part 1: primary test metric at the validation-selected layer for HyPLM-Flash and the main encoders of similar size. Bold marks the best value in each row.

Task	Metric	ESM-2 650M	ESM-1b	ESM C 600M	E1-600M	AMPLIFY-350M	DPLM-650M	**HyPLM-Flash**
Zero-shot and likelihood								
ProteinGym zero-shot	Spearman ρ	0.415	0.389	0.406	**0.420**	0.363	0.410	0.409
NLL, held-out UniRef panel	NLL	1.820	1.862	1.720	1.811	1.743	1.961	**1.698**
NLL, recent PDB chains	NLL	1.591	1.502	1.478	1.560	1.155	1.796	**1.148**
Structure								
Secondary structure (CB513)	Q3 accuracy	0.844	0.814	0.851	0.880	0.797	0.842	**0.882**
Contacts, ProteinNet	P@$L/5$, long range	0.392	0.410	0.434	0.645	0.210	0.394	**0.650**
Contacts, CASP14	P@$L/5$, long range	0.366	0.322	0.385	0.537	0.198	0.367	**0.556**
Contacts, CASP15	P@$L/5$, long range	0.391	0.404	0.440	0.566	0.244	0.394	**0.574**
Remote homology (fold)	accuracy	**0.373**	0.362	0.345	**0.373**	0.286	0.370	0.322
Function, localization and interaction								
EC, SaProt split	$F_{\max}$	0.888	**0.889**	0.886	0.881	0.865	0.882	0.885
EC, DeepFRI split	$F_{\max}$	0.874	**0.878**	0.876	0.871	0.861	0.874	0.874
GO-MF	$F_{\max}$	**0.667**	0.659	0.665	0.661	0.631	0.663	0.661
GO-BP	$F_{\max}$	0.496	0.485	**0.499**	0.497	0.467	0.496	0.492
GO-CC	$F_{\max}$	0.535	0.520	0.533	0.519	0.524	0.530	**0.538**
EC classification (DGEB)	macro F1	0.665	0.673	0.701	**0.715**	0.669	0.682	0.687
Arch retrieval (DGEB)	MAP@5	0.120	**0.206**	0.035	0.042	0.057	0.194	0.086
Localization, 10-class	accuracy	0.806	0.793	0.808	**0.839**	0.729	0.806	0.794
Localization, binary	accuracy	0.907	0.922	0.912	0.910	0.891	**0.922**	0.906
Metal-ion binding	accuracy	0.725	0.690	0.678	0.710	**0.737**	0.728	0.695
HumanPPI	accuracy	0.856	0.867	0.856	**0.900**	0.811	0.833	0.817
Fitness and property regression								
AAV two-vs-many	Spearman ρ	0.518	0.544	0.551	0.583	0.574	0.569	**0.650**
AAV low-vs-high	Spearman ρ	0.279	0.280	0.296	0.295	0.254	0.286	**0.296**
GB1 two-vs-rest	Spearman ρ	0.552	0.615	0.640	**0.660**	0.497	0.584	0.628
GB1 three-vs-rest	Spearman ρ	0.801	0.807	0.762	0.765	0.774	0.777	**0.808**
GB1 low-vs-high	Spearman ρ	0.377	0.417	0.395	**0.474**	0.273	0.404	0.469
Meltome human cell	Spearman ρ	0.696	0.698	0.697	**0.720**	0.665	0.684	0.684
Meltome mixed	Spearman ρ	0.633	0.593	0.670	**0.690**	0.585	0.623	0.672
Amylase one-to-many	Spearman ρ	0.640	0.653	0.681	0.579	0.592	0.674	**0.696**
Amylase by mutation count	Spearman ρ	**0.575**	0.503	0.512	0.539	0.298	0.502	0.553
Hydrolase to P06241	Spearman ρ	0.172	−0.224	0.203	−0.295	0.212	**0.380**	−0.192
Fluorescence	Spearman ρ	0.664	0.659	0.663	0.660	0.660	0.668	**0.669**
Stability	Spearman ρ	0.639	0.596	0.706	0.712	0.579	**0.746**	0.687
Thermostability	Spearman ρ	0.697	0.699	0.702	**0.721**	0.664	0.684	0.684

Table 7: All tasks, part 2: other encoders. – not reported (Appendix A).

Task	Metric	ESM3-open	ProtBERT	ProtT5-XL	Ankh-base	Ankh3-large	xTrimoPGLM-1B
Zero-shot and likelihood							
ProteinGym zero-shot	Spearman ρ	**0.409**	0.343	–	–	–	–
NLL, held-out UniRef panel	NLL	1.920	**1.754**	–	–	–	–
NLL, recent PDB chains	NLL	1.828	**1.105**	–	–	–	–
Structure							
Secondary structure (CB513)	Q3 accuracy	**0.896**	0.785	0.844	0.847	0.810	0.799
Contacts, ProteinNet	P@$L/5$, long range	**0.568**	0.229	0.566	0.546	0.383	0.321
Contacts, CASP14	P@$L/5$, long range	0.479	0.200	**0.499**	0.467	0.307	0.302
Contacts, CASP15	P@$L/5$, long range	0.521	0.250	0.521	**0.523**	0.369	0.352
Remote homology (fold)	accuracy	**0.408**	0.270	0.326	0.362	0.288	0.364
Function, localization and interaction							
EC, SaProt split	$F_{\max}$	0.880	0.861	**0.886**	0.881	0.868	0.877
EC, DeepFRI split	$F_{\max}$	0.878	0.860	**0.880**	0.870	0.854	0.872
GO-MF	$F_{\max}$	**0.664**	0.622	0.660	0.659	0.640	0.649
GO-BP	$F_{\max}$	0.493	0.454	**0.495**	0.484	0.476	0.484
GO-CC	$F_{\max}$	0.508	0.529	**0.533**	0.502	0.512	0.524
EC classification (DGEB)	macro F1	**0.681**	0.589	0.671	0.680	0.645	0.502
Arch retrieval (DGEB)	MAP@5	0.277	0.072	0.193	**0.311**	0.270	0.124
Localization, 10-class	accuracy	0.750	0.739	**0.831**	0.792	0.764	0.739
Localization, binary	accuracy	0.900	0.894	**0.919**	0.906	0.918	0.906
Metal-ion binding	accuracy	0.659	0.707	0.741	**0.746**	0.692	0.719
HumanPPI	accuracy	0.822	0.833	**0.889**	0.833	0.839	0.839
Fitness and property regression							
AAV two-vs-many	Spearman ρ	**0.551**	0.409	0.533	0.539	0.496	0.322
AAV low-vs-high	Spearman ρ	0.287	0.231	0.253	**0.296**	0.267	0.270
GB1 two-vs-rest	Spearman ρ	0.519	0.507	0.414	**0.617**	0.299	0.592
GB1 three-vs-rest	Spearman ρ	0.766	0.765	0.781	0.796	**0.814**	0.770
GB1 low-vs-high	Spearman ρ	0.455	0.326	0.402	0.386	**0.471**	0.461
Meltome human cell	Spearman ρ	0.682	0.660	0.682	0.669	0.668	**0.698**
Meltome mixed	Spearman ρ	0.573	0.621	**0.661**	0.527	0.545	0.578
Amylase one-to-many	Spearman ρ	0.618	0.544	**0.682**	0.628	0.653	0.656
Amylase by mutation count	Spearman ρ	0.524	0.484	0.398	**0.548**	0.353	0.359
Hydrolase to P06241	Spearman ρ	**0.527**	0.246	0.268	0.096	−0.128	0.230
Fluorescence	Spearman ρ	0.669	0.641	0.663	0.642	0.669	**0.673**
Stability	Spearman ρ	0.758	0.516	0.586	0.649	**0.759**	0.660
Thermostability	Spearman ρ	0.679	0.661	0.680	0.670	0.684	**0.696**