E1: Retrieval-Augmented Protein Encoder Models

Sarthak Jain [1] **Joel Beazer** [1] **Jeffrey A. Ruffolo** [1] **Aadyot Bhatnagar** [1] **Ali Madani** [1]

Abstract

Large language models trained on natural proteins learn powerful representations of protein sequences that are useful for downstream understanding and prediction tasks. Because they are only exposed to individual protein sequences during pretraining without any additional contextual information, conventional protein language models suffer from parameter inefficiencies in learning, baked-in phylogenetic biases, and functional performance issues at larger scales. To address these challenges, we introduce E1, a family of retrieval-augmented protein language models that explicitly condition on homologous sequences. By integrating retrieved evolutionary context through block-causal multi-sequence attention, E1 captures both general and family-specific constraints without fine-tuning. We train E1 models on four trillion tokens from the Profluent Protein Atlas and achieve state-of-the-art performance across zero-shot fitness and unsupervised contact-map prediction benchmarks – surpassing alternative sequence-only models. Performance scales with model size from 150M to 600M parameters, and E1 can be used flexibly in single-sequence or retrieval-augmented inference mode for fitness prediction, variant ranking, and embeddings for structural tasks. To encourage open science and further advances in retrieval-augmented protein language models, we release three models for free research and commercial use at `https://github.com/Profluent-AI/E1`.

1. Introduction

Proteins are central to biological processes—including catalysis, transport, immunity, and gene regulation—and underpin applications in pharmaceuticals, agriculture, and biotechnology. Protein engineering aims to design sequences with desired functions, but mapping sequence to function remains challenging, with many approaches still relying on random mutagenesis and high-throughput screening.

Protein language models (PLMs) provide a data-driven framework for modeling sequence–structure–function relationships. Trained on large protein databases, they capture evolutionary patterns shaped by natural selection (Ruffolo & Madani, 2024). Single-sequence models such as ESM-2/C (ESM Team, 2024; Lin et al., 2023; Elnaggar et al., 2023; Brandes et al., 2022; Elnaggar et al., 2021) produce likelihoods correlated with function and learn representations encoding structural and functional information (Meier et al., 2021; Lin et al., 2023).

However, single-sequence PLMs have key limitations: they compress evolutionary context into model parameters, limiting performance on underrepresented families, and may reflect dataset biases rather than true functional constraints (Zhang et al., 2024b; Gordon et al., 2024; Ding & Steinhardt, 2024; Bhatnagar et al., 2025). Fine-tuning can help but is often costly and impractical for data-scarce settings.

To overcome these limitations, recent approaches have incorporated explicit evolutionary context through retrieval augmentation. Retrieval-augmented PLMs (RA-PLMs) enhance standard single-sequence models by providing homologous sequences during training and inference. This allows the model to leverage evolutionary context directly. Notable examples include the MSA Transformer (Rao et al., 2021), which uses multiple sequence alignments as context, and PoET (Truong Jr & Bepler, 2023), which employs alignment-free concatenations of homologous sequences.

In this work, we introduce **`E1`**, a new family of retrieval-augmented protein encoder models trained with a masked language modeling objective. We leverage a large-scale Protein Atlas (Bhatnagar et al., 2025) and introduce targeted architectural and training innovations that yield more performant RA-PLMs. **`E1`** achieves state-of-the-art performance among models trained exclusively on sequence data. On the Protein Gym benchmark for zero-shot fitness prediction, **`E1`** models outperform the ESM family (ESM Team, 2024; Lin et al., 2023) in single-sequence mode and surpass other retrieval-based models, including PoET (Truong Jr & Bepler, 2023) and MSA Pairformer (Akiyama et al., 2025),

[1]Profluent Bio, Emeryville, CA, USA. Correspondence to: Sarthak Jain <sjain@profluent.bio>, Ali Madani <ali@profluent.bio>.

Proceedings of the ICML 2026 3rd Workshop on Multi-modal Foundation Models and Large Language Models for Life Sciences, Seoul, Korea. 2026.

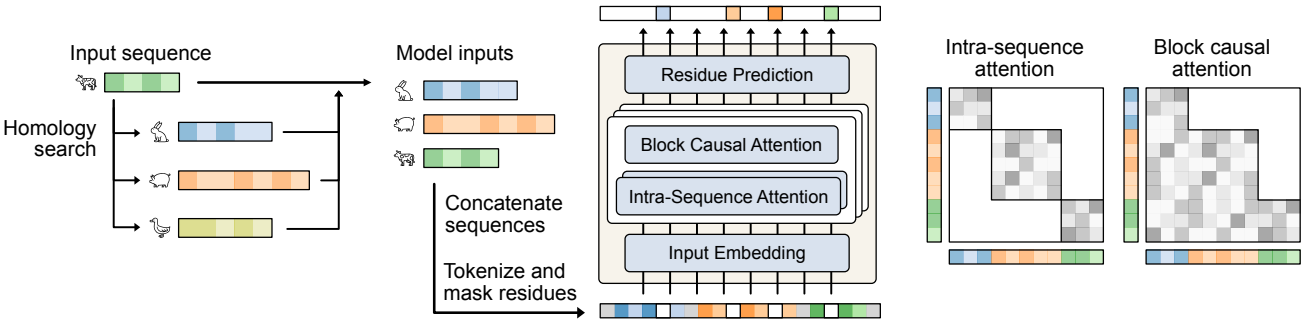

Figure 1. **E1 Architecture.** The **E1** model can take in homologous sequences in addition to an input query sequence. The homologous sequences are prepended to the query sequence to construct a multi-sequence input to the model. **E1** alternates between intra-sequence and block-causal attention, enabling it to build internal representations based on residues within the same protein sequence as well as residues in preceding homologous sequences within the concatenated multi-sequence input.

when augmented with homologs. **E1** also achieves superior performance in unsupervised contact-map prediction, again outperforming the ESM family in single-sequence mode and showing substantial additional gains with retrieval. We also observe that the performance of our models scales with the number of parameters. We release three **E1** variants – 150M, 300M, and 600M parameters – freely for research and commercial use, enabling immediate application to tasks such as fitness prediction, structure prediction, and representation learning.

2. Model

2.1. Architecture

E1 is a family of retrieval-augmented protein encoder models trained with bidirectional attention and a masked language modeling objective. In contrast to standard protein encoder models like ESM-2 (Lin et al., 2023), these models leverage sequence homologs as part of their inference context to generate better representations for a given sequence of interest, allowing for in-context learning. Note that we do not require the homologous sequences to be aligned with each other, in contrast to models like MSA Transformer (Rao et al., 2021) and MSA Pairformer (Akiyama et al., 2025). To test whether the model performance scales with number of parameters, we trained three different sizes of **E1** models: 150M, 300M, and 600M parameters.

The model takes as input a sequence of protein sequences (for example, `MLFH,MIIVR,MFHK`) with each individual sequence wrapped in special tokens (`<bos>1MLFH2<eos> <bos>1MIIVR2<eos> <bos>1MFHK2<eos>`) to mark the start and end of the sequence. Embeddings of these tokens are then passed to the model. Each token in the same protein sequence also shares a sequence ID, which is then embedded and supplied to the model to distinguish between different protein sequences within a multi-sequence instance. We allow up to 512 individual sequences within a single multi-sequence instance. **E1** model family is implemented using a standard Transformer-based architecture (Vaswani et al., 2017; Devlin et al., 2019), augmented with a block causal attention mechanism that enables residues in different homologous sequences to attend to one another. For efficiency, this global attention is not applied in every layer. Instead, we adopt an alternating attention architecture (Warner et al., 2025): global block-causal attention is used every three layers, while all other layers use intra-sequence attention, where residues attend only to other residues within the same protein sequence.

We use standard Rotary Position Embedding (RoPE) (Su et al., 2024) to encode positional information. For layers using intra-sequence attention, each protein sequence restarts position IDs at one. In contrast, for global-attention layers, the position ID corresponds to the absolute position of the token within the full concatenated multi-sequence input.

2.2. Training

The **E1** family of models was trained using a standard masked language modeling objective (Devlin et al., 2019), in which input tokens are randomly selected and replaced with noisy variations. A language modeling head (a single hidden layer MLP) is then applied on top of the final-layer token representations to predict the probability of the true amino acid at each selected position. During training, we linearly decreased the noise fraction (the fraction of tokens replaced in the input) from 25% to 15% for the first 250 billion tokens; after that, it remained fixed at 15%. We followed the standard BERT masking policy: 80% of selected tokens were replaced with a special mask token, 10% were replaced with a random amino acid, and the remaining 10% were left unchanged. All three **E1** models were trained for 4 trillion tokens (batch size = 2^{20} tokens) using a WSD learning rate schedule (Hu et al., 2024) and Stable AdamW optimizer (Wortsman et al., 2023).

2.3. Training Data Construction

To construct multi-sequence instances for training, we adopt the strategy introduced by the PoET model (Truong Jr & Bepler, 2023). We used sets of homologous sequences derived from the PPA-1 (Bhatnagar et al., 2025) and UniRef Version 2411 (Suzek et al., 2015) datasets. Both PPA-1 and UniRef are clustered at multiple sequence identity thresholds, including at 50% and 90% identity. For each 50% ID cluster representative, we search it against all other 50% ID cluster representatives in the respective datasets using Diamond (Buchfink et al., 2021), returning a set of possible homologs. To construct a training instance, we first randomly sample one of these homolog sets (with probability inversely proportional to the size of the set) and then replace each 50% ID cluster representative with a randomly picked sequence from the associated 50% ID sequence cluster (weighted inversely by the size of its 90% ID subcluster). Finally, we subset the resulting sequences to ensure that the concatenated multi-sequence instance remains within a prescribed length budget.

We employed a curriculum learning strategy where we gradually increased the total length and number of sequences in a multi-sequence instance: from 8192 to 32768 and from 2 to 512 respectively. This enabled the model to achieve state of the art performance in both single sequence mode (where no homologous sequences are passed during inference) and retrieval-augmented mode. During training, we exclusively trained on instances from PPA-1 for the first 1.5 trillion tokens. Thereafter, we mixed in instances from UniRef in a 60:40 ratio for the remainder of the training duration.

3. Results

3.1. `E1` models enable state of the art zero-shot substitution effect prediction

Protein language models have been shown to be effective zero-shot fitness predictors for local mutational landscapes. In addition, prior work (Rao et al., 2021; Truong Jr & Bepler, 2023; 2025; Tan et al., 2024; Sun et al., 2024; Zhang et al., 2024a) has shown that addition of evolutionarily related sequences (either unaligned or in the form of an MSA) during inference can improve the model's performance. In this section, we use the 217 Deep Mutational Scan substitution assays from the ProteinGym (v1.3) benchmark (Notin et al., 2023) to evaluate the performance of **`E1`** models in both single-sequence and retrieval-augmented modes. We use the masked marginal method (Meier et al., 2021) to compute scores for each variant of the wildtype protein sequence and evaluate performance using Spearman correlation and the normalized discounted cumulative gain (NDCG) metric against ground truth fitness values. The latter metric measures the ability of the model to rank high fitness sequences first and is more practically relevant for protein design tasks.

Table 1. Average Spearman correlation and NDCG@10 between model-predicted scores and Protein Gym experimental fitness values.

Model Name	Spearman Corr. Average	NDCG@10 Average
Inference with query sequence only		
ESM2-150M	0.387	0.729
ESM2-650M	0.414	0.747
ESM2-3B	0.406	**0.755**
ESMC-300M	0.406	0.746
ESMC-600M	0.405	0.746
`E1` 150M	0.401	0.744
`E1` 300M	0.416	0.748
`E1` 600M	**0.420**	0.749
Inference with Homologous Sequences / MSA		
MSA Pairformer	0.45	—
PoET	0.470	0.784
`E1` 150M	0.473	0.785
`E1` 300M	0.475	0.787
`E1` 600M	**0.477**	**0.788**

Sampling homologs for inference. For evaluation in retrieval-augmented mode, we follow the PoET strategy (Truong Jr & Bepler, 2023) and prepend the masked variants of the wildtype sequence with homologous sequences sampled from ColabFold derived MSAs (Mirdita et al., 2022) constructed using Uniref100 v2104. Homologs are sampled with weights inversely proportional to the number of their neighbors (sequences in the MSA that are at least 80% identical to them)and are additionally constrained to satisfy a specified maximum similarity to the wildtype sequence. We ensemble 15 prompts corresponding to 3 different total-token-length budgets and 5 different maximum query-similarity thresholds ($\{6144, 12288, 24576\} \times \{1.0, 0.95, 0.9, 0.7, 0.5\}$).

Results. In Table 1, we observe that **`E1`** models outperform all ESM-2 and ESMC family models in single-sequence mode at comparable model sizes, indicating that **`E1`** can be used as a drop-in replacement for existing single-sequence encoder models without loss of performance. When evaluated with homologs at inference time, the **`E1`** models substantially outperform corresponding single-sequence metrics and achieve state of the art performance relative to similar publicly available models, i.e., models that only take homologous sequences as additional context during inference, like MSA Pairformer and PoET[1]. In Table 4 (Appendix A), we further observe that switching from single-sequence to retrieval-augmented mode yields consistent improvements

[1]The metrics for MSA Pairformer are taken from the original paper, while PoET, ESM-2, and ESMC are sourced from the Protein Gym public leaderboard

for assays with low and medium MSA depth. On average, the larger **E1** models also tend to perform better, indicating continued benefits of scaling up retrieval-augmented PLMs.

3.2. Unsupervised contact map prediction benefits from homologous sequences during inference

Unsupervised contact map prediction can be used as an efficient proxy to test whether the model has learned to encode information about the 3D structures of proteins during pretraining. In this section, we compare the performance of **E1** with publicly available models on the long-range contact prediction task for protein sequences from CAMEO (Haas et al., 2018; Robin et al., 2021) and CASP15 (Kryshtafovych et al., 2023) targets. We use the Categorical Jacobian approach (Zhang et al., 2024b) to assess the model's internal knowledge of residue–residue contacts in an architecture-agnostic manner and report precision-at-L (the percentage of top-L predicted contacts that are correct). We define a residue pair as being in contact if their Cβ-Cβ distance is $< 8\mathring{A}$, and we define long-range contact as contact between residues separated by at least 24 positions in sequence space.

Table 2. Unsupervised contact map prediction performance as measured by Precision@L for long range contacts.

Model Name	Long-range Precision@L CAMEO	CASP15
Query Sequence Only		
ESM2-150M	0.348	0.272
ESM2-650M	0.423	0.342
ESM2-3B	0.434	0.339
ESMC-300M	0.425	0.342
E1 150M	0.466	0.387
E1 300M	0.493	0.401
E1 600M	**0.512**	**0.425**
Query Sequence + Homologs/MSA		
MSA Pairformer	0.489	0.428
E1 150M	0.510	0.406
E1 300M	0.526	0.415
E1 600M	**0.541**	**0.436**

We also evaluate whether the model can exploit additional information from homologous sequences during inference to improve contact-prediction performance. Homologs are sampled using the same procedure described in the previous section, with MSAs generated by ColabFold from the UniRef dataset. We fix the context length to 8192 and the maximum query similarity to 0.95 and use a single prompt for evaluation.

Results. We observe from Table 2 that **E1** models outperform the ESM family of models at all scales when tested in single-sequence mode. Moreover, we see consistent gains in performance when including homologous sequences during inference, indicating that the model is able to leverage in-context evolutionary information to identify putative 3D contacts in a protein. Finally, we provide some illustrative examples from the CAMEO dataset in Figure 2 where retrieval augmentation yields markedly improved contact-map predictions relative to single-sequence inference.

4. Discussion

We introduced **E1**, a family of retrieval-augmented protein encoder models that can leverage unaligned evolutionarily related sequences at inference time to achieve superior performance. **E1** achieves state-of-the-art performance among publicly available models on variant-effect prediction (Protein Gym) and unsupervised contact-map prediction (CAMEO and CASP15), both in single-sequence mode and when augmented with homologs. We release three **E1** variants – 150M, 300M, and 600M – that are available for free for research and commercial use.

While these results demonstrate the benefits of retrieval augmentation, several questions remain about the model's behavior. In particular, it is unclear how much **E1** relies on information stored in pretrained weights versus signals from retrieved homologs. Unlike alignment-based models such as *MSA Transformer* (Rao et al., 2021), which use structured attention over aligned sequences, **E1** allows unrestricted cross-sequence attention within the multi-sequence input. This raises the question of whether the model implicitly recovers alignment-like correspondences or instead exploits broader contextual information beyond traditional MSAs.

We observe scaling trends with increasing model size on zero-shot tasks, though our study is limited to models with up to 600M parameters. More broadly, we focus on sequence-only models to isolate the effects of retrieval augmentation; incorporating structural information during pretraining may further improve performance and efficiency (Sun et al., 2024; Hayes et al., 2025; Su et al., 2023; Truong Jr & Bepler, 2025). Finally, an open question is whether conditioning on homologs with specific properties can steer the model toward desired regions of the fitness landscape, enabling more targeted protein design.

Overall, the **E1** family of models demonstrates the continued value of research in improving protein language models and provides a new foundational tool for AI-driven protein design that advances both predictive performance and practical utility for a large class of protein design workflows.

References

Akiyama, Y., Zhang, Z., Mirdita, M., Steinegger, M., and Ovchinnikov, S. Scaling down protein language modeling with msa pairformer. *bioRxiv*, pp. 2025–08, 2025.

Bhatnagar, A., Jain, S., Beazer, J., Curran, S. C., Hoffnagle,

A. M., Ching, K. S., Martyn, M., Nayfach, S., Ruffolo, J. A., and Madani, A. Scaling unlocks broader generation and deeper functional understanding of proteins. *bioRxiv*, pp. 2025–04, 2025.

Brandes, N., Ofer, D., Peleg, Y., Rappoport, N., and Linial, M. Proteinbert: a universal deep-learning model of protein sequence and function. *Bioinformatics*, 38(8):2102–2110, 2022.

Buchfink, B., Reuter, K., and Drost, H.-G. Sensitive protein alignments at tree-of-life scale using diamond. *Nature methods*, 18(4):366–368, 2021.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. In *North American Chapter of the Association for Computational Linguistics*, 2019. URL https://api.semanticscholar.org/CorpusID:52967399.

Ding, F. and Steinhardt, J. 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.

Elnaggar, A., Heinzinger, M., Dallago, C., Rehawi, G., Wang, Y., Jones, L., Gibbs, T., Feher, T., Angerer, C., Steinegger, M., et al. Prottrans: Toward understanding the language of life through self-supervised learning. *IEEE transactions on pattern analysis and machine intelligence*, 44(10):7112–7127, 2021.

Elnaggar, A., Essam, H., Salah-Eldin, W., Moustafa, W., Elkerdawy, M., Rochereau, C., and Rost, B. Ankh: Optimized protein language model unlocks general-purpose modelling. *bioRxiv*, pp. 2023–01, 2023.

ESM Team. Esm cambrian: Revealing the mysteries of proteins with unsupervised learning, 2024. URL https://evolutionaryscale.ai/blog/esm-cambrian.

Gordon, C., Lu, A. X., and Abbeel, P. Protein language model fitness is a matter of preference. *bioRxiv*, pp. 2024–10, 2024.

Haas, J., Barbato, A., Behringer, D., Studer, G., Roth, S., Bertoni, M., Mostaguir, K., Gumienny, R., and Schwede, T. Continuous automated model evaluation (cameo) complementing the critical assessment of structure prediction in casp12. *Proteins: Structure, Function, and Bioinformatics*, 86:387–398, 2018.

Hayes, T., Rao, R., Akin, H., Sofroniew, N. J., Oktay, D., Lin, Z., Verkuil, R., Tran, V. Q., Deaton, J., Wiggert, M., et al. Simulating 500 million years of evolution with a language model. *Science*, 387(6736):850–858, 2025.

Hu, S., Tu, Y., Han, X., He, C., Cui, G., Long, X., Zheng, Z., Fang, Y., Huang, Y., Zhao, W., et al. Minicpm: Unveiling the potential of small language models with scalable training strategies. *arXiv preprint arXiv:2404.06395*, 2024.

Kryshtafovych, A., Schwede, T., Topf, M., Fidelis, K., and Moult, J. Critical assessment of methods of protein structure prediction (casp)—round xv. *Proteins: Structure, Function, and Bioinformatics*, 91(12): 1539–1549, 2023. doi: https://doi.org/10.1002/prot.26617. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/prot.26617.

Lin, Z., Akin, H., Rao, R., Hie, B., Zhu, Z., Lu, W., Smetanin, N., Verkuil, R., Kabeli, O., Shmueli, Y., et al. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science*, 379(6637): 1123–1130, 2023.

Meier, J., Rao, R., Verkuil, R., Liu, J., Sercu, T., and Rives, A. Language models enable zero-shot prediction of the effects of mutations on protein function. *Advances in Neural Information Processing Systems*, 34:29287–29303, 2021.

Mirdita, M., Schütze, K., Moriwaki, Y., Heo, L., Ovchinnikov, S., and Steinegger, M. Colabfold: making protein folding accessible to all. *Nature methods*, 19(6):679–682, 2022.

Notin, P., Kollasch, A., Ritter, D., Van Niekerk, L., Paul, S., Spinner, H., Rollins, N., Shaw, A., Orenbuch, R., Weitzman, R., et al. Proteingym: Large-scale benchmarks for protein fitness prediction and design. *Advances in Neural Information Processing Systems*, 36:64331–64379, 2023.

Rao, R. M., Liu, J., Verkuil, R., Meier, J., Canny, J., Abbeel, P., Sercu, T., and Rives, A. Msa transformer. In *International conference on machine learning*, pp. 8844–8856. PMLR, 2021.

Robin, X., Haas, J., Gumienny, R., Smolinski, A., Tauriello, G., and Schwede, T. Continuous automated model evaluation (cameo)—perspectives on the future of fully automated evaluation of structure prediction methods. *Proteins: Structure, Function, and Bioinformatics*, 89 (12):1977–1986, 2021.

Ruffolo, J. A. and Madani, A. Designing proteins with language models. *Nature Biotechnology*, 42(2):200–202, 2024.

Su, J., Han, C., Zhou, Y., Shan, J., Zhou, X., and Yuan, F. Saprot: Protein language modeling with structure-aware vocabulary. *BioRxiv*, pp. 2023–10, 2023.

Su, J., Ahmed, M., Lu, Y., Pan, S., Bo, W., and Liu, Y. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063, 2024.

Sun, N., Zou, S., Tao, T., Mahbub, S., Li, D., Zhuang, Y., Wang, H., Cheng, X., Song, L., and Xing, E. P. Mixture of experts enable efficient and effective protein understanding and design. *bioRxiv*, pp. 2024–11, 2024.

Suzek, B. E., Wang, Y., Huang, H., McGarvey, P. B., Wu, C. H., and Consortium, U. Uniref clusters: a comprehensive and scalable alternative for improving sequence similarity searches. *Bioinformatics*, 31(6):926–932, 2015.

Tan, Y., Wang, R., Wu, B., Hong, L., and Zhou, B. Retrieval-enhanced mutation mastery: Augmenting zero-shot prediction of protein language model. *arXiv preprint arXiv:2410.21127*, 2024.

Truong Jr, T. and Bepler, T. Poet: A generative model of protein families as sequences-of-sequences. *Advances in Neural Information Processing Systems*, 36:77379–77415, 2023.

Truong Jr, T. F. and Bepler, T. Understanding protein function with a multimodal retrieval-augmented foundation model. *arXiv preprint arXiv:2508.04724*, 2025.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. Attention is all you need. In *Advances in neural information processing systems*, pp. 5998–6008, 2017.

Warner, B., Chaffin, A., Clavié, B., Weller, O., Hallström, O., Taghadouini, S., Gallagher, A., Biswas, R., Ladhak, F., Aarsen, T., et al. Smarter, better, faster, longer: A modern bidirectional encoder for fast, memory efficient, and long context finetuning and inference. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 2526–2547, 2025.

Wortsman, M., Dettmers, T., Zettlemoyer, L., Morcos, A., Farhadi, A., and Schmidt, L. Stable and low-precision training for large-scale vision-language models. *Advances in Neural Information Processing Systems*, 36: 10271–10298, 2023.

Zhang, Z., Notin, P., Huang, Y., Lozano, A. C., Chenthamarakshan, V., Marks, D., Das, P., and Tang, J. Multi-scale representation learning for protein fitness prediction. *Advances in Neural Information Processing Systems*, 37: 101456–101473, 2024a.

Zhang, Z., Wayment-Steele, H. K., Brixi, G., Wang, H., Kern, D., and Ovchinnikov, S. Protein language models learn evolutionary statistics of interacting sequence motifs. *Proceedings of the National Academy of Sciences*, 121(45):e2406285121, 2024b.

A. Additional Protein Gym Results

Table 3. Average Spearman correlation and NDCG@10 between model-predicted scores and Protein Gym experimental fitness values.

Model Name	Spearman Correlation						NDCG@10
	Average	Activity	Binding	Expression	Organismal Fitness	Stability	Average
Inference with query sequence only							
ESM2-150M	0.387	0.391	0.326	0.402	0.305	0.51	0.729
ESM2-650M	0.414	0.425	**0.337**	0.415	0.368	0.523	0.747
ESM2-3B	0.406	0.417	0.321	0.403	**0.378**	0.509	**0.755**
ESMC-300M	0.406	0.423	0.315	0.408	0.36	0.526	0.746
ESMC-600M	0.405	0.423	0.294	0.42	0.362	0.528	0.746
E1 150M	0.401	0.426	0.325	0.420	0.304	0.532	0.744
E1 300M	0.416	**0.438**	0.332	0.430	0.346	0.537	0.748
E1 600M	**0.420**	0.415	0.330	**0.441**	0.366	**0.548**	0.749
Inference with Homologous Sequences / MSA in-context							
MSA Pairformer	0.45	0.49	0.35	0.44	0.46	0.51	—
PoET	0.470	0.494	0.396	0.466	0.475	0.519	0.784
E1 150M	0.473	0.498	0.408	0.468	0.477	0.514	0.785
E1 300M	0.475	**0.501**	**0.410**	0.468	0.474	0.523	0.787
E1 600M	**0.477**	**0.501**	0.404	**0.469**	**0.478**	**0.532**	**0.788**

Table 4. Average Spearman correlation between model-predicted scores and Protein Gym experimental fitness values broken down by Taxon and MSA Depth.

Model Name	Spearman Correlation by Taxon				Spearman Correlation by MSA Depth		
	Human	Other Eukaryote	Prokaryote	Virus	Low	Medium	High
Inference with query sequence only							
ESM2-150M	0.45	0.475	0.398	0.157	0.319	0.359	0.494
ESM2-650M	0.457	0.486	0.458	0.261	0.338	0.409	0.513
ESM2-3B	0.442	0.477	0.458	0.294	0.336	0.423	0.485
ESMC-300M	0.468	0.481	0.441	0.242	0.337	0.399	0.520
ESMC-600M	0.462	0.481	0.459	0.241	0.331	0.407	0.515
E1 150M	0.455	0.515	0.413	0.188	0.342	0.373	0.514
E1 300M	0.466	0.513	0.444	0.238	0.367	0.396	0.524
E1 600M	0.475	0.482	0.472	0.254	0.342	0.419	0.523
Inference with Homologous Sequences / MSA in-context							
PoET	0.482	0.541	0.464	0.491	0.478	0.478	0.510
E1 150M	0.482	0.527	0.476	0.494	0.476	0.477	0.515
E1 300M	0.485	0.534	0.478	0.490	0.471	0.480	0.520
E1 600M	0.487	0.537	0.488	0.500	0.478	0.485	0.525

B. Example Contact Prediction from E1 Models

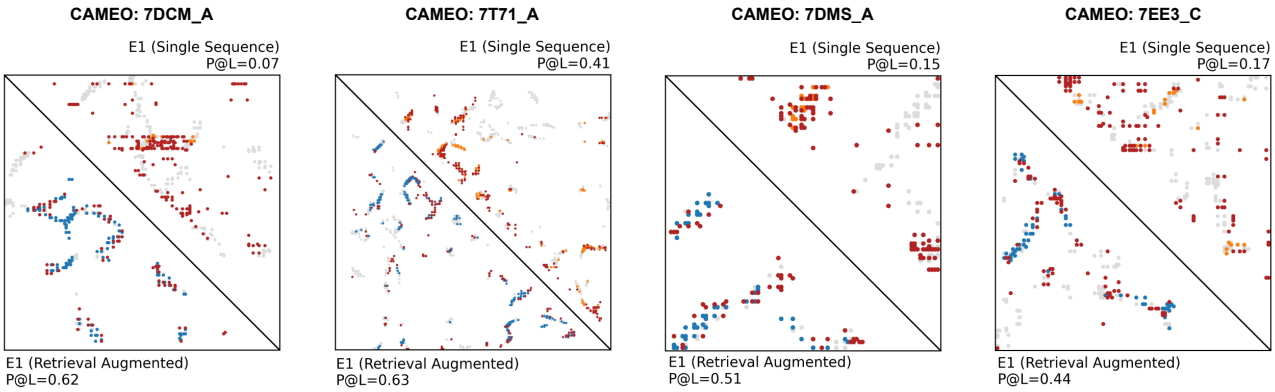

Figure 2. Examples from CAMEO dataset where retrieval augmentation helps **E1** identify contact it may have mispredicted when used in single sequence mode. Here, gray points are ground truth contacts, blue/orange points are correctly predicted contacts in retrieval-augmented/single-sequence mode, respectively, and red points are false positives.