Retrieved Sequence Augmentation for Protein Representation Learning

Chang Ma [1] **Haiteng Zhao** [2] **Lin Zheng** [1] **Jiayi Xin** [1] **Qintong Li** [1] **Lijun Wu** [3] **Zhihong Deng** [2] **Yang Lu** [4] **Qi Liu** [1] **Lingpeng Kong** [1]

Abstract

Protein language models have excelled in a variety of tasks, ranging from structure prediction to protein engineering. However, proteins are highly diverse in functions and structures, and current state-of-the-art models including the latest version of AlphaFold rely on Multiple Sequence Alignments (MSA) to feed in the evolutionary knowledge. Despite their success, heavy computational overheads, as well as the de novo and orphan proteins remain great challenges in protein representation learning. In this work, we show that MSA-augmented models inherently belong to retrieval-augmented methods. Motivated by this finding, we introduce **R**etrieved **S**equence **A**ugmentation (RSA) for protein representation learning without additional alignment or pre-processing. RSA links query protein sequences to a set of sequences with similar structures or properties in the database and combines these sequences for downstream prediction. We show that protein language models benefit from the retrieval enhancement on both structure prediction and property prediction tasks, with a 5% improvement on MSA Transformer on average while being 373× faster. In addition, we show that our model can transfer to new protein domains better and outperforms MSA Transformer on de novo protein prediction. Our study fills a much-encountered gap in protein prediction and brings us a step closer to demystifying the domain knowledge needed to understand protein sequences. Code is available on https://github.com/HKUNLP/RSA.

[1]Department of Computer Science, The University of Hong Kong [2]School of Intelligence Science and Technology, Peking University [3]Microsoft Research Asia [4]Department of Computer Science, University of Waterloo. Correspondence to: Chang Ma <changma@connect.hku.hk>, Lingpeng Kong <lpk@cs.hku.hk>.

1. Introduction

Proteins are the basic yet intricate building blocks of life, performing a vast array of functions within organisms, including catalyzing metabolic reactions, DNA replication, responding to stimuli, providing structure to cells, and transporting molecules from one location to another (Garrett & Grisham, 2016). Central to the enigma of these building blocks is the complex knowledge of protein relationships in their sequences, structures, and functions, which is a consequence of the interplay between physics and evolution (Sadowski & Jones, 2009). Experimental and theoretical efforts have been made to unveil the structures and functions of emergent proteins (Korendovych & DeGrado, 2020; Anishchenko et al., 2021), yet few methods can keep pace with the rapid accumulation of sequences (Roy et al., 2010).

Recently, protein language models (Rives et al., 2019; Lin et al., 2022; Elnaggar et al., 2021; Jumper et al., 2021) have achieved remarkable progress in predicting protein functions and structures from sequences. Protein language models create a distribution of amino acids that matches the co-occurrence probability in their natural state, thereby capturing structural and evolutionary knowledge. In these approaches, all protein knowledge is implicitly stored in the parameters, and the quality of the language model distribution is highly dependent on pre-training and parameter scale. For example, ESM-2 (Lin et al., 2022) shows that evolutionary depth saturates at lower model scales, and scaling up to a model size of billions is inevitable for protein modeling. To this end, we study enhancing the prediction of language models with a simple retrieval-based augmentation.

Previous work (Khandelwal et al., 2019; Goyal et al., 2022; Guu et al., 2020b; Wang et al., 2022) in natural language processing and machine learning has demonstrated that introducing related input sequences can effectively introduce domain knowledge without excessive backbone parameter size. In protein learning, a similar approach Multiple Sequence Alignment (MSA) has been adopted to introduce evolutionary knowledge into models by augmenting input with aligned homologous sequences. MSA has improved deep learning performance on various models (Rao et al., 2021; Jumper et al., 2021; Marks et al., 2011; Hong et al., 2022), yet its success is often attributed to the alignment

process that highlights co-evolution – especially the alignment process that is central to direct-coupling analysis methods (Morcos et al., 2011; Marks et al., 2011; Kamisetty et al., 2013). The most common practice for constructing MSA (Remmert et al., 2012; Altschul & Koonin, 1998; Johnson et al., 2010) is to build a Hidden Markov Model (HMM) profile for the entire sequence space of databases and then iteratively search for homologous sequences. Despite efforts to accelerate MSA construction (Remmert et al., 2012; Deorowicz et al., 2016; Hauser et al., 2016), this process is notoriously slow – it takes HHblits (Remmert et al., 2012) 10 seconds to perform a single iteration search on Pfam with 64 CPUs – and requires pre-computing of a HMM profile.

These considerations motivate us to rethink the role of MSA as a retrieval-based augmentation. Viewing MSA as a retrieval-augmentation method, it can be decomposed into two processes: retrieval and alignment. As shown in Figure 1, the speed bottleneck of MSA is the alignment time, which is constrained by a quadratic complexity of $O(LD)$ (Remmert et al., 2012), where D is the database size, and L is the protein length. Meanwhile, dense retrievers can be accelerated and use only a 100th of the time MSA needs to align a sequence (Hong et al., 2021; Johnson et al., 2019b). Moreover, the language of proteins encodes not only evolutionary knowledge but also other sources of information including structural and functional properties (Xia et al., 2009; O'Sullivan et al., 2004). Multiple sources of knowledge can be used to aid protein understanding when evolutionary knowledge is not available for orphan proteins and de novo (designed) proteins (Perdigão et al., 2015; Stefani, 2004; Anishchenko et al., 2021). Residue alignment imitates the mutation process in proteins, but empirically, present large language models have the potential to directly capture the evolutionary relationship between sequences without alignment information (Riesselman et al., 2019).

In light of these bottlenecks, We propose a simple yet effective **R**etrieved **S**equence **A**ugmentation (RSA) method as a general framework for augmenting protein sequences with related sequences from an unlabeled database. Specifically, RSA uses a pre-trained dense sequence retriever to retrieve protein sequences that are similar to the query sequence both in terms of homology as well as structure. These sequences are learned together with original input to help the model cover external knowledge and transfer to new domains. Extensive experiments on six tasks, including secondary structure prediction, contact prediction, homology prediction, stability prediction, subcellular localization, and protein-protein interaction demonstrate the effectiveness of our model. In addition, RSA overcomes the speed limit of MSA methods by directly inputting a batch of retrieved sequences into protein language models without performing the alignment process. Our main contributions are:

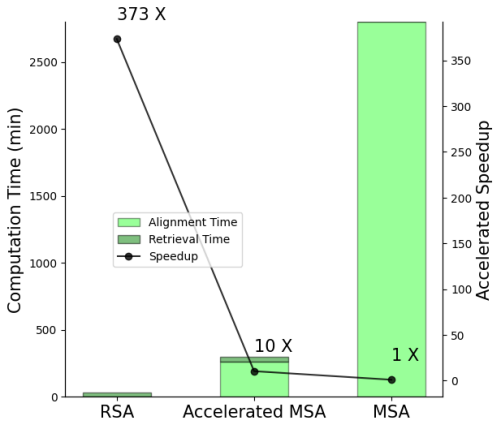

Figure 1. Illustration of speed up by RSA retrieval compared to MSA on secondary structure prediction dataset with 8678 sequences. Accelerated MSA refers to the MSA Transformer with MSA sequences retrieved by our RSA retriever.

- Employing probabilistic analysis, we develop a unified framework that uses retrieval knowledge to enhance protein language models. Our theory along with our experiments strikes two novel perspectives: (1) MSA-augmented methods are essentially retrieval-augmented language models. Their performance can be explained by the injection of evolutionary knowledge. (2) The $O(N^2)$ complex alignment process is less necessary for deep protein language models.

- We show that pre-trained dense retrievers can be faster and perform well in extracting homologous sequences and structurally similar sequences.

- We leverage the retrieval augmentation framework to develop a new, fast method RSA. Unlike previous methods that combine protein language models with external knowledge, our method performs retrieval on-the-fly and requires no additional pre-training. We show that our model performs better than or competitively with previous SOTAs. The result promises new opportunities in using retrieval augmentation as a new paradigm in protein learning. Code and data are available in the supplementary material.

2. Related Work

Retrieval-Augmented Language Models The scaling laws of language models indicate that scaling up model size and training data are central to better performance (Kaplan et al., 2020). However, larger language models are expensive to pre-train and may even be computationally heavy in inference. Retrieval-augmented language models (Guu et al., 2020a; He et al., 2021a; Borgeaud et al., 2022) can achieve

comparable performance on smaller models and are computationally more efficient by injecting external knowledge. Our RSA method is motivated by retrieval-augmented language models (Guu et al., 2020a; He et al., 2021a), though we specifically focus on injecting protein knowledge and adapt the model for token-level tasks and better efficiency.

Protein Language Models To model and further understand the protein sequence data, language models are introduced to train on mass data (Heinzinger et al., 2019; Alley et al., 2019). Large scale pre-training enables language models to learn structural and evolutionary knowledge (Elnaggar et al., 2021; Jumper et al., 2021; Lin et al., 2022). Despite these successes, many important applications still require MSAs and other external knowledge (Rao et al., 2021; Jumper et al., 2021; He et al., 2021b; Zhang et al., 2021; Ju et al., 2021; Rao et al., 2020). MSAs have been shown effective in improving representation learning, despite being extremely slow and costly in computation. Hu et al. (2022) and Hong et al. (2021) use dense retrieval to accelerate multiple sequence augmentation, while still dependent on alignment procedures. Recent work (Fang et al., 2022; Lin et al., 2022; Wu et al., 2022; Chowdhury et al., 2022) explores MSA-free language models though additional pre-training is involved. We take this step further to investigate retrieval-augmented protein language models that finds a balance between large scale pre-training and external knowledge.

3. Problem Statement and Notations

The task of protein representation learning is to learn embeddings of protein sequences that can be transferred to downstream tasks with finetuning. For a protein x with L amino acids, it can be denoted as $x = [o_1, o_2, ...o_L]$, where each token o_i denotes one of the 25 essential amino acids. We implement the embedding functions using BERT-style Transformer encoder $\text{Embed}(x) = [h_1, h_2, ...h_L]^T$, where $h_i \in \mathbb{R}^d$ is a d-dimensional token representation for o_i. For token property prediction (i.e., secondary structure prediction), pairwise prediction (i.e., contact prediction), and sequence property prediction (i.e., protein engineering) tasks, the probabilities are obtained through pooling operations defined below:

$$
\begin{aligned}
&p(y_{\text{Token}}|o_i) = \text{FFN}(h_i),\\
&p(y_{\text{Pairwise}}|o_i, o_j) = \text{FFN}([h_i; h_j]),\\
&p(y_{\text{Sequence}}|x) = \text{FFN}(\text{Mean}([h_1, h_2, ...h_L]).
\end{aligned}
$$

4. MSA Transformer as a Retrieval Augmentation Method

In this section, we introduce a unified probabilistic framework to connect the MSA-based models with retrieval augmentations. We also offer a new holistic view on understanding these models, that is the retrieved protein sequences enhance the performance of pre-trained protein models by providing evolutionary knowledge in a similar way as the MSA sequences do.

Inspired by Guu et al. (2020a) and the probabilistic form of MSA Transformer, we propose a general framework, *protein retrieval augmentation*, that aims to unify several state-of-the-art evolution augmentation methods. Specifically, we consider these methods as learning a downstream predictor $p(y|x)$ based on an aggregation of homologous protein representations $R_{1...N}$. From the view of retrieval, $p(y|x)$ is decomposed into two steps: *retrieve* and *predict*. For a given input x, the retrieve step first finds possibly helpful protein sequence r from a sequence corpus $\mathcal{R}$ and then predict the output y conditioning on this retrieved sequence. We treat r as a latent variable and in practice, we approximately marginalized it out with top-N retrieved sequences:

$$
p(y|x) = \sum_{r\in\mathcal{R}} p(y|x, r)p(r|x) \approx \sum_{n=1}^{N} p(y|x, r_n)p(r_n|x). \tag{1}
$$

The probability $p(r|x)$ denotes the possibility that r is sampled from the retriever given x. Intuitively it measures the similarity between the two sequences r and x. This framework also applies to the MSA-based augmentation methods. We explain in detail using a state-of-the-art MSA-augmentation model *MSA Transformer* (Rao et al., 2021) as an example. In MSA Transformer, the layers calculate self-attention both row-wise and column-wise. Column-wise attention is defined as follows, given W_Q, W_K, W_V, W_O as the parameters in a typical attention function:

$$
R_s(i) = \sum_{n=1}^{N} \sigma(\frac{R_s(i)W_Q(R_n(i)W_K)^T}{N\sqrt{d}})R_n(i)W_V W_O, \tag{2}
$$

where $R_n(i)$ denotes the i-th token representation of the n-th MSA sequence after performing the row-wise attention. Note that in MSA input, the first sequence r_1 is defined as the original sequence x. Then for a token prediction task, we define the i-th position output as y and the predicted distribution $p(y|x)$ can be expressed as:

$$
\begin{aligned}
p(y|x) &= \sum_{n=1}^{N} \sigma(\frac{R_1W_Q(R_nW_K)^T}{N\sqrt{d}})(R_nW_V W_O W_y)\\
&= \sum_{n=1}^{N} p(y|x, r_n)\lambda_n = \sum_{n=1}^{N} p(y|x, r_n)p(r_n|x),
\end{aligned} \tag{3}
$$

where $\lambda_n = \sigma(\frac{R_1(i)W_Q(R_n(i)W_K)^T}{N\sqrt{d}})$ is the weighting norm that represents the similarity of retrieved sequence r_n and original sequence x; $p(y|x, r_n)$ is a predictor that maps the row-attention representation of r_n and x to label.

Table 1. Protein Retrieval Augmentation methods decomposed along a different axis. We formulate the aggregation function in the sequence classification setting and use a feed-forward neural network FFN$(\cdot)$ to map representations to logits. The proposed variants vary in design axis from the existing methods. †Note that MSA Transformer performs the aggregation in each layer of axial attention.

Method	Retriever Form	Alignment Form	Weight λ_n	Aggregation Function
		Existing Methods		
Potts Model	MSA	Aligned	—	—
Co-evolution Aggregator	MSA	Aligned	$\frac{1}{N}$	FFN$(\sum_{n=1}^{N} R_n(i)\lambda_n)$
MSA Transformer	MSA	Aligned	$\sigma(\frac{XW_Q(R_nW_K)^T}{N\sqrt{d}})$	FFN$(\sum_{n=1}^{N} R_n(i)\lambda_n)^\dagger$
		Proposed Variants		
Unaligned MSA Augmentation	MSA	Not Aligned	$\sigma(-\|\|X - R_n\|\|_2)$	$\sum_{n=1}^{N}$ FNN$(R_n(i))\lambda_n$
Accelerated MSA Transformer	Dense Retrieval	Aligned	$\sigma(\frac{XW_Q(R_nW_K)^T}{N\sqrt{d}})$	FFN$(\sum_{n=1}^{N} R_n(i)\lambda_n)$
Retrieval Sequence Augmentation	Dense Retrieval	Not Aligned	$\sigma(-\|\|X - R_n\|\|_2)$	$\sum_{n=1}^{N}$ FFN(Embed$(x; r_n)$)λ_n

Eq.3 gives a retrieval-augmentation view of MSA Transformer that essentially retrieves homologous sequences with multiple sequence alignment and aggregates representations of homologous sequences with regard to their sequence similarity. Taking one step further, we define a set of design dimensions to characterize the retrieving and aggregation processes. We detail the design dimensions below and illustrate how popular models (Appendix B) and our proposed methods (§5) fall along them in Table 1. These design choices includes:

- **Retriever Form** indicates the retriever type used. Multiple Sequence Alignment is a discrete retrieval method that uses E-value thresholds (Ye et al., 2006) to find homologous sequences. Dense retrieval (Johnson et al., 2019b) has been introduced to accelerate discrete sequence retrieval. The method represents the database with dense vectors and retrieves the sequences that have top-k vector similarity with the query.
- **Alignment Form** indicates whether retrieved sequences are aligned, as illustrated in Appendix Figure 6.
- **Weight Form** is the aggregation weight of homologous sequences, as the $p(r_n|x)$ in Eq. 3. Here we denote this weight as λ_n. Traditionally, aggregation methods consider the similarity of different homologous sequences to be the same and use average weighting. MSA Transformer also use a weighted pooling method though the weights of λ_n use global attention and are dependent on all homologous sequences.
- **Aggregation Function** is how the representations of homologous sequences are aggregated to the original sequence to form downstream prediction, as in $p(y|x, r)$. For example, considering the sequence classification problem, a fully connected layer maps representations to logits. MSA Transformer first aggregates the representations R_n and then maps the aggregated representation to logits y, and the retrieval augmentation probabilistic form first maps each representation to logits $p(y|x, r_n)$ and then linearly weight the logits with λ_n in Eq. 3.

Our discussion and formulation so far reach the conclusion that MSA augmentation methods intrinsically use the retrieval augmentation approach. This highlights the potential of RSA to replace MSA Augmentations as a computationally effective and more flexible method.

However, MSA-based methods claim a few advantages: the *alignment* process can help the model capture column-wise residue evolution; and the *MSA Retriever* uses a discrete, token-wise search criterion that ensures all retrieved sequences are homology. We propose two novel variants to help verify these claims.

Unaligned MSA Augmentation. MSA modeling traditionally depends on the structured alignment between sequences to learn evolutionary information. However, deep models have the potential to learn patterns from unaligned sequences. Riesselman et al. (2019) shows that the mutation effect can be learned from unaligned sequences using autoregressive models. Therefore, we first introduce this variant that uses the homologous sequences from MSA to augment representations without alignment.

Accelerated MSA Transformer. This variant explores substituting the discrete retrieval process in MSA with a dense retriever. We use the K-nearest neighbor search to find the homologous sequences. We still align the sequences before input into MSA Transformer. We introduce this variant to find if MSA builder has an advantage over our pre-trained dense retriever in finding related sequences.

An empirical study of the performance of these models can be found in Subsection 6.6.

5. Retrieval Sequence Augmentations

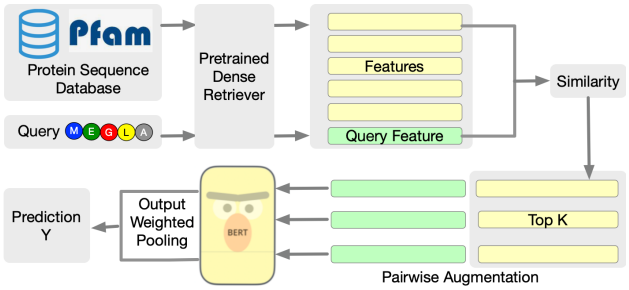

Figure 2. A brief overview of the proposed RSA protein encoding framework. Based on a query protein, RSA first retrieves related protein data from the database based on the top K similar features encoded by a pretrained retrieval model. Then we augment the query protein into pairs with each retrieved data and feed them into the protein model for protein tasks.

Table 2. Recall and Precision for retrieving top 100 protein sequences with ESM1b embeddings. In dataset Pfam and SCOPe, we test whether retrieved proteins are of the same Family, Superfamily, or Fold as query protein, and report the recall and precision.

Retrieval Task (Top 100)	Type	Recall	Precision
Pfam - Family	Homology	100	90.42
SCOPe - Fold	Structural	100	65.98
SCOPe - Superfamily	Structural	100	46.00
SCOPe - Family	Structural	100	24.71

Existing knowledge augmentation methods for protein representation learning are either designed for a specific task or require cumbersome data preprocessing. Motivated by the potential of pre-trained retrievers to identify proteins that are homologous or geometric similar, we propose a pipeline, RSA (**R**etrieval **S**equence **A**ugmentation), to directly augment protein models on-the-fly. Our model implementation follows the *retrieve-then-predict* framework in Eq. 1. We elaborate on the model architecture implementations in Subsection 5.1 and describe model training in Subsection 5.2.

5.1. Model Architectures

The RSA model comprises of a neural sequence retriever $p(r|x)$, and a protein model that combines both original input and retrieved sequence to obtain prediction $p(y|x, r)$.

5.1.1. RSA Retriever

The retriever is defined as finding the sequences that are semantically close to the query. Denote retriever model as G which encode protein sequence and output embeddings.

$$p(r|x) = \frac{\exp f(x,r)}{\sum_{r' \in \mathcal{R}} \exp f(x,r')}, \qquad f(x,r) = -||G(x) - G(r)||_2 \tag{4}$$

The similarity score $f(x, r)$ is defined as the negative L2 distance between the embedding of the two sequences. The distribution is the softmax distribution over similarity scores.

For protein retrieval, we aim to retrieve protein sequences that have similar structures or are homologous to the query sequence. Motivated by the k-nearest neighbor retrieval experiment with ESM-1b (Rives et al., 2019) pre-trained embeddings (as shown in Table 2 and Figure 4), we implement the embedding functions using a 34-layer ESM-1b encoder. We obtain sequence embeddings by performing average pooling over token embeddings. Note that finding the most similar proteins from a large-scale sequence database is computationally heavy. To accelerate retrieval, we use Faiss indexing (Johnson et al., 2019a), which uses clustering of dense vectors and quantization to allow efficient similarity search at a massive scale.

5.1.2. RSA Encoder

Retrieval Augmented Protein Encoder Given a sequence x and a retrieved sequence r with length L and M respectively, the protein encoder combines x and r for prediction $p(y|x, r)$. To make our model applicable to any protein learning task, we need to augment both sequence-level representation and token-level representation. To achieve this, we concatenate the two sequences before input into the transformer encoder, which uses self-attention to aggregate global information from the retrieved sequence r into each token representation.

$$A = \sigma(\frac{(H_{[x;r]}W^Q)(H_{[x;r]}W^K)^T}{\sqrt{d}}), A = [A_x; A_r]$$
$$Attn(H_{[x;r]}) = (A_x H_x W^V + A_r H_r W^V)W^O$$

where $H_{[x;r]} = [h_1^x, h_2^x, ..., h_L^x, h_1^r ... h_M^r]^T$ denotes the input embedding of original and retrieved sequences. The output token representation h_i automatically learns to select and combine the representation of retrieved tokens. This can also be considered a soft version of MSA alignment. After computing for each pair of (x, r), we aggregate them by weight $p(r|x)$ defined in Eq. 4.

5.2. RSA Training

Training For downstream finetuning, we maximize $p(y|x)$ by performing training on the retrieval augmented protein encoder. We freeze the retriever parameters during training. For a query sequence with N retrieved proteins, the computation cost is N times the original model, $O(NL^2)$ for a

Table 3. Main Results for vanilla protein representation learning methods, knowledge-augmented baselines and our proposed RSA method. Note that *italized* result is reported by corresponding related work. The last column reports average result on all six tasks. For MSA Transformer and RSA, we all use 16 sequences (N=16) for augmentation. For Gremlin Potts model, we use the full MSA.

Method	Pretrain	Knowledge Pretrain	Knowledge Injection	SSP	Contact	Homology	Stability	Loc	PPI	Avg
Transformer	×	×	×	0.384	0.274	0.101	0.422	0.541	0.616	0.345
LSTM	×	×	×	*0.596*	*0.263*	*0.181*	*0.591*	*0.629*	*0.638*	0.404
RSA (Transformer backbone)	×	×	✓	0.541	0.332	0.346	0.602	0.591	0.700	0.518
ESM-1b	✓	×	×	*0.716*	*0.458*	*0.978*	*0.695*	*0.781*	*0.782*	0.668
ProtBERT	✓	×	×	0.691	0.556	0.528	0.651	0.771	0.688	0.579
MSA Transformer (MSA N=1)	✓	✓	×	0.594	0.397	0.880	0.767	0.668	0.633	0.592
Gremlin (Balakrishnan et al., 2011)	×	×	✓	—	0.507	—	—	—	—	—
MSA Transformer	✓	✓	✓	0.654	0.618	0.958	**0.796**	0.694	0.751	0.672
OntoProtein (Zhang et al., 2022)	✓	×	✓	*0.68*	*0.40*	0.96	*0.75*	—	—	—
PMLM (He et al., 2021b)	✓	✓	×	***0.728***	***0.717***	*0.946*	—	—	—	—
RSA (ProtBERT backbone)	✓	×	✓	0.691	**0.717**	**0.987**	0.778	**0.795**	**0.827**	**0.723**

transformer encoder layer, which is more efficient than the MSA Transformer with a $O(NL^2)+O(N^2L)$ computation cost. Also, the retrieval is performed on the fly.

6. Experiments

6.1. General Setup

Downstream tasks In order to evaluate the performance of our trained model, six datasets are introduced, namely secondary structure prediction, contact prediction, remote homology prediction, subcellular localization prediction, stability prediction, and protein-protein interaction. Please refer to Appendix Table 9 for more statistics of the datasets. The train-eval-test splits follow TAPE benchmark (Rao et al., 2019) for the first four tasks and PEER benchmark (Xu et al., 2022) for subcellular localization and protein-protein interaction. The introduction to datasets is in Appendix C.1.

Retriever and MSA Setup Limited by available computation resources, we build a database on Pfam (El-Gebali et al., 2018) sequences, which covers 77.2% of the UniProtKB (Apweiler et al., 2004) database and reaches the evolutionary scale. We generate ESM-1b pre-trained representations of 44 million sequences from Pfam-A and use Faiss (Johnson et al., 2019b) to build the retrieval index. For a fair comparison, the MSA datasets are also built on the Pfam database. We use HHblits (Remmert et al., 2012) to extract MSA. The details are shown in Appendix C.2.

Baselines We apply our retrieval method to both pre-trained and randomly initialized language models. Following Rao et al. (2019) and Rao et al. (2021), we compare our model with vanilla protein representation models, including LSTM(Liu, 2017), Transformers(Vaswani et al., 2017) and pre-trained models ESM-1b(Rives et al., 2019), ProtBERT(Elnaggar et al., 2020). We also compare with state-of-the-art knowledge-augmentation models: Potts Model(Balakrishnan et al., 2011), MSA Transformer(Rao et al., 2021) that inject evolutionary knowledge through MSA, OntoProtein(Zhang et al., 2022) that uses gene ontology knowledge graph to augment protein representations and PMLM(He et al., 2021b) that uses pair-wise pretraining to improve co-evolution awareness. We use the reported results of LSTM from Zhang et al. (2021); Xu et al. (2022).

Training and Evaluation Our RSA model is applicable to any global-aware encoders. To demonstrate RSA as a general method, we perform experiments both with a shallow transformer encoder, and a large pre-trained ProtBERT encoder. The Transformer model has 512 dimensions and 6 layers. All self-reported models use the same truncation strategy and perform parameter searches on the learning rate, warm-up rate, seed, and batch size. For evaluation, we choose the best-performing model on the validation set and perform prediction on the test set.

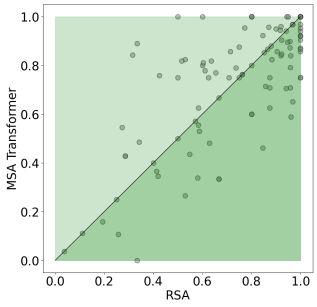

Figure 3. Contact Prediction of RSA and MSA Transformer on De Novo Proteins. We plot samples that RSA have better predictions under the diagonal line.

6.2. Main Results

We show the result for downstream tasks in Table 3, including models with/without pretraining, and with/without knowledge augmentations. We form the following conclusion: **Retrieval Sequence Augmentations perform on par with or even better than other knowledge-augmented methods without additional pre-training.** The last two blocks compare our method with previous augmentation methods. Our method outperforms MSA Transformer on average by 5% and performs on par with PMLM on structure and evolution prediction tasks. Notably, both MSA Transformer and PMLM perform additional pre-training with augmentations, while our method uses no additional pre-training. From the results, we can see that RSA combined transformer model also improves by 10% than other shallow models, demonstrating the effectiveness of our augmentation to both shallow models and pre-trained models.

Table 4. The table shows remote homology prediction performance with increasing domain gaps: Family, Superfamily and Fold.

Method	Family	Superfam	Fold
Transformer	0.101	0.518	0.078
MSA Transformer (no MSA)	0.880	0.278	0.206
ProtBERT	0.528	0.192	0.170
MSA Transformer	0.958	0.503	0.235
Accelerated MSA Transformer	0.945	0.406	0.227
RSA (ProtBERT backbone)	**0.987**	**0.677**	**0.267**

6.3. Retrieval Augmentation for Domain Adaptation

We investigate the model's transfer performance in domains with distribution shifts. We train our model on the Remote Homology dataset, and test it on three testsets with increasing domain gaps: proteins that are within the same Family, Superfam, and Fold as the training set respectively. The results are in Table 4. It is pertinent to note that MSA transformer's performance decreases dramatically when the gap between the domains increases. Our model surpasses MSA Transformer by a large margin on shifted domains, especially from 0.5032 to 0.6770 on Superfam. Our model proves to be more reliable for domain shifts, illustrating that retrieval facilitates the transfer across domains.

Furthermore, we test our model on 108 out-of-domain De Novo proteins for the contact prediction task. De Novo proteins are synthesized by humans and have a different distribution from natural proteins. It can be seen in Figure 3 that, in addition to surpassing MSA transformer on average precision by 1%, RSA also exceeds MSA transformer on 63.8% of data, demonstrating that RSA is more capable of locating augmentations for out-of-distribution proteins. We also test our model on the secondary structure task with new domain data, as shown in Appendix (Table 8 and Figure 7). The results also show that our model surpasses MSA Transformer in transferring to unseen domains.

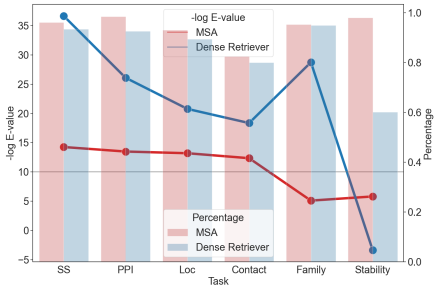

Figure 4. Plot of the -log(E-values) of MSA and Dense Retriever obtained sequences on the test sets for six tasks. E-values of both methods are obtained with HHblits(Remmert et al., 2012). Sequences with -log E-value >10 are high-quality homologous sequences. We also show with bar plots the percentage of sequences in the test sets that have homologous sequences.

Table 5. Results for MSA Transformer and Unaligned MSA Augmentation on Homology and Stability task. Both models use MSA as inputs, but Unaligned MSA Augmentation unaligns MSA and augments the model by concatenating MSA sequence to the input.

Methods	Homology	Stability
MSA Transformer	0.958	**0.796**
Unaligned MSA Augmentation	0.973	0.749
RSA	**0.987**	0.778

6.4. Retrieval Speed

A severe speed bottleneck limits the use of previous MSA-based methods. In this part, we compare the computation time of RSA with MSA and an accelerated version of MSA as introduced in Section 4. As shown in Figure 1, alignment time cost is much more intense than retrieval time. Even after reducing the alignment database size to 500, accelerated MSA still need 270 min to build MSA. At the same time RSA only uses dense retrieval, and is accelerated 373 times. Note that with extensive search, MSA can find *all* available alignments in a database. However, this would be less beneficial to deep protein language models as the memory limit only suffices a few dozens of retrieved sequences.

6.5. Retrieved Protein Interpretability

The previous retrieval-augmented language models rely on a dense retriever to retrieve knowledge-relevant documents. However, it remains indistinct what constitutes knowledge for protein understanding and how retrieved sequences can be used for improving protein representations. In this section, we take a close look at the retrieved protein sequences to examine their homology and geometric properties.

Dense Retrievers Find Homologous Sequences. One type of knowledge distinct to the protein domain is sequence

homology, which infers knowledge on shared ancestry between proteins in evolution. Homologous sequences are more likely to share functions or similar structures. We analyze whether retrieved sequences are homologous.

As illustrated in Figure 4 (right axis), across all six datasets, our dense retriever retrieved a high percentage of homologous proteins that can be aligned to the original protein sequence, comparable to traditional HMM-based MSA retrievers. We additionally plot each dataset's negative log E-values distribution in Figure 4. Accordingly, pre-trained protein models can be used directly as dense retrieval of homologous sequences.

Table 6. Results for MSA Transformer and Accelerated MSA Transformer on downstream tasks. Accelerated MSA Transformer uses MSA built from dense retrieval sequences.

Methods	MSA Transformer	Accelerated MSA Transformer	RSA
SSP	0.654	0.634	**0.691**
Contact	0.618	0.608	**0.717**
Homology	0.958	0.945	**0.987**
Stability	**0.796**	0.767	0.778
Loc	0.694	0.682	**0.795**
PPI	0.751	0.679	**0.827**

RSA Retriever Find Structurally Similar Protein Protein structures are also central to protein functions and properties. In this section, we analyze whether retrieved sequences are structurally similar. In Figure 5, we plot the TM scores between the RSA retrieved protein and the origin protein on ProteinNet (AlQuraishi, 2019) test set. Using ESMFold[1], we obtain the 3D structures of the top 5 retrieved proteins and then calculate the TM score between these proteins and the query protein. Most of the retrieved proteins exceed the 0.2 criteria, which indicates structural similarity, and about half are above the 0.5 criteria, which indicates high quality. Accordingly, this indicates that the dense retrieval algorithm is capable of finding proteins with structural knowledge.

6.6. Ablation Study

Ablation on Retriever: Unaligned MSA Augmentation. We ablate RSA retriever by using MSA retrieved proteins as augmentations to our model, denoted as Unaligned MSA Augmentation. The results are in Table 5. As the result shows, Unaligned MSA Augmentation performs worse than our RSA model, especially on the Stability dataset, where the performance drops from 0.778 to 0.7443. It thus confirms the ability of our dense retriever to provide more abundant knowledge for protein models.

Ablation on Retriever: Ablation on Retrieval Number Our study examines the effect of injected knowledge quantity for RSA and all retrieval baselines. The results are listed

[1]https://esmatlas.com/resources?action=fold

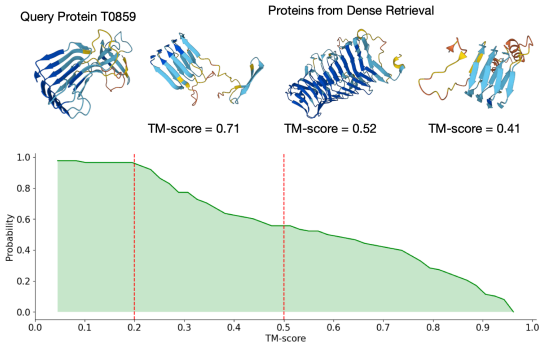

Figure 5. Plot of the cumulative distribution of TM-scores for proteins from dense retrieval. The value at a shows the probability that TM-score is larger than a. We also give a visual example of retrieved protein to illustrate similar structures.

in Table 7. We select the Contact dataset because all baseline models are implemented on this dataset. RSA and all baselines perform consistently better as the retrieval number increases. Also, our model outperforms all baseline models for all augmentation numbers.

Table 7. The performance of retrieval augmentation models w.r.t. the number of retrieved sequences on contact prediction.

Methods	N=1	N=4	N=8	N=16	N=32	N= full
Potts Model	—	0.412	0.471	0.479	0.480	0.507
MSA Transformer	0.397	0.579	0.560	0.618	0.669	—
Accelerated MSA Transformer	0.397	0.524	0.538	0.608	0.654	—
RSA	**0.556**	**0.595**	**0.615**	**0.717**	**0.719**	—

Ablation on aggregation: We compare RSA with Accelerated MSA Transformer to evaluate whether our aggregation method is beneficial for learning protein representations. Note that only part of the retrieved sequences that satisfy homologous sequence criteria are selected and utilized during alignment. As shown in Table 6, the performance of the Accelerated MSA Transformer drops a lot compared to RSA. In contrast to MSA type aggregation, which is restricted by token alignment, our aggregation is more flexible and can accommodate proteins with variant knowledge.

Is MSA retriever necessary? Table 6 illustrates that Accelerated MSA Transformer performs near to MSA Transformer (MSA N=16) for most datasets, except for Stability and PPI on which our retriever failed to find enough homologous sequences, as Figure 4 demonstrates. Our retriever is therefore capable of finding homologous sequences for most tasks and is able to replace the MSA retriever.

Is MSA alignment necessary? To support that MSA alignment is not necessary, we compare Unaligned MSA Augmentation to the original MSA transformer. As revealed by the results in Table 5. Unaligned MSA Augmentation performs close to the MSA transformer. This confirms our declaration that self-attention is capable of integrating pro-

tein sequences into representations.

7. Conclusions and Future Work

In this paper, we introduce a simple yet effective method to enhance protein representation learning. We demonstrate RSA as a fast yet high-performing method that has the potential to replace MSA-based methods in most scenarios. For future work, we hope to further scale up our RSA method and apply it to 3D folding tasks.

References

Alley, E. C., Khimulya, G., Biswas, S., AlQuraishi, M., and Church, G. M. Unified rational protein engineering with sequence-based deep representation learning. *Nature methods*, 16(12):1315–1322, 2019.

Almagro Armenteros, J. J., Sønderby, C. K., Sønderby, S. K., Nielsen, H., and Winther, O. Deeploc: prediction of protein subcellular localization using deep learning. *Bioinformatics*, 33(21):3387–3395, 2017.

AlQuraishi, M. Proteinnet: a standardized data set for machine learning of protein structure. *BMC bioinformatics*, 20(1):1–10, 2019.

Altschul, S. F. and Koonin, E. V. Iterated profile searches with psi-blast—a tool for discovery in protein databases. *Trends in biochemical sciences*, 23(11):444–447, 1998.

Anishchenko, I., Pellock, S. J., Chidyausiku, T. M., Ramelot, T. A., Ovchinnikov, S., Hao, J., Bafna, K., Norn, C., Kang, A., Bera, A. K., et al. De novo protein design by deep network hallucination. *Nature*, 600(7889):547–552, 2021.

Apweiler, R., Bairoch, A., Wu, C. H., Barker, W. C., Boeckmann, B., Ferro, S., Gasteiger, E., Huang, H., Lopez, R., Magrane, M., et al. Uniprot: the universal protein knowledgebase. *Nucleic acids research*, 32(suppl_1): D115–D119, 2004.

Balakrishnan, S., Kamisetty, H., Carbonell, J. G., Lee, S.-I., and Langmead, C. J. Learning generative models for protein fold families. *Proteins: Structure, Function, and Bioinformatics*, 79(4):1061–1078, 2011.

Bank, P. D. Rcsb pdb. 2022, 2022.

Borgeaud, S., Mensch, A., Hoffmann, J., Cai, T., Rutherford, E., Millican, K., Van Den Driessche, G. B., Lespiau, J.-B., Damoc, B., Clark, A., et al. Improving language models by retrieving from trillions of tokens. In *International conference on machine learning*, pp. 2206–2240. PMLR, 2022.

Chowdhury, R., Bouatta, N., Biswas, S., Floristean, C., Kharkar, A., Roy, K., Rochereau, C., Ahdritz, G., Zhang, J., Church, G. M., et al. Single-sequence protein structure prediction using a language model and deep learning. *Nature Biotechnology*, 40(11):1617–1623, 2022.

Deorowicz, S., Debudaj-Grabysz, A., and Gudyś, A. Famsa: Fast and accurate multiple sequence alignment of huge protein families. *Scientific reports*, 6(1):1–13, 2016.

El-Gebali, S., Mistry, J., Bateman, A., Eddy, S. R., Luciani, A., Potter, S. C., Qureshi, M., Richardson, L. J., Salazar, G. A., Smart, A., Sonnhammer, E. L., Hirsh, L., Paladin, L., Piovesan, D., Tosatto, S. C., and Finn, R. D. The Pfam protein families database in 2019. *Nucleic Acids Research*, 47(D1):D427–D432, 10 2018. ISSN 0305-1048. doi: 10.1093/nar/gky995. URL https://doi.org/10.1093/nar/gky995.

Elnaggar, A., Heinzinger, M., Dallago, C., Rihawi, G., Wang, Y., Jones, L., Gibbs, T., Feher, T., Angerer, C., Steinegger, M., et al. Prottrans: towards cracking the language of life's code through self-supervised deep learning and high performance computing. *arXiv preprint arXiv:2007.06225*, 2020.

Elnaggar, A., Heinzinger, M., Dallago, C., Rehawi, G., Wang, Y., Jones, L., Gibbs, T., Feher, T., Angerer, C., Steinegger, M., Bhowmik, D., and Rost, B. Prottrans: Towards cracking the language of life's code through self-supervised learning. *bioRxiv*, 2021.

Fang, X., Wang, F., Liu, L., He, J., Lin, D., Xiang, Y., Zhang, X., Wu, H., Li, H., and Song, L. Helixfold-single: Msa-free protein structure prediction by using protein language model as an alternative. *arXiv preprint arXiv:2207.13921*, 2022.

Garrett, R. H. and Grisham, C. M. *Biochemistry*. Cengage Learning, 2016.

Goyal, A., Friesen, A., Banino, A., Weber, T., Ke, N. R., Badia, A. P., Guez, A., Mirza, M., Humphreys, P. C., Konyushova, K., et al. Retrieval-augmented reinforcement learning. In *International Conference on Machine Learning*, pp. 7740–7765. PMLR, 2022.

Guu, K., Lee, K., Tung, Z., Pasupat, P., and Chang, M. Retrieval augmented language model pre-training. In *International Conference on Machine Learning*, pp. 3929–3938. PMLR, 2020a.

Guu, K., Lee, K., Tung, Z., Pasupat, P., and Chang, M.-W. Realm: Retrieval-augmented language model pre-training. *international conference on machine learning*, 2020b.

Hauser, M., Steinegger, M., and Söding, J. Mmseqs software suite for fast and deep clustering and searching of large protein sequence sets. *Bioinformatics*, 32(9):1323–1330, 2016.

He, J., Neubig, G., and Berg-Kirkpatrick, T. Efficient nearest neighbor language models. *arXiv preprint arXiv:2109.04212*, 2021a.

He, L., Zhang, S., Wu, L., Xia, H., Ju, F., Zhang, H., Liu, S., Xia, Y., Zhu, J., Deng, P., et al. Pre-training co-evolutionary protein representation via a pairwise masked language model. *arXiv preprint arXiv:2110.15527*, 2021b.

Heinzinger, M., Elnaggar, A., Wang, Y., Dallago, C., Nechaev, D., Matthes, F., and Rost, B. Modeling aspects of the language of life through transfer-learning protein sequences. *BMC bioinformatics*, 20(1):1–17, 2019.

Hong, L., Sun, S., Zheng, L., Tan, Q., and Li, Y. fastmsa: Accelerating multiple sequence alignment with dense retrieval on protein language. *bioRxiv*, 2021.

Hong, Y., Song, J., Ko, J., Lee, J., and Shin, W.-H. S-pred: protein structural property prediction using msa transformer. *Scientific reports*, 12(1):1–11, 2022.

Hou, J., Adhikari, B., and Cheng, J. Deepsf: deep convolutional neural network for mapping protein sequences to folds. *Bioinformatics*, 34(8):1295–1303, 2018.

Hu, M., Yuan, F., Yang, K. K., Ju, F., Su, J., Wang, H., Yang, F., and Ding, Q. Exploring evolution-aware &-free protein language models as protein function predictors. In Oh, A. H., Agarwal, A., Belgrave, D., and Cho, K. (eds.), *Advances in Neural Information Processing Systems*, 2022. URL https://openreview.net/forum?id=U8k0QaBgXS.

Johnson, J., Douze, M., and Jégou, H. Billion-scale similarity search with GPUs. *IEEE Transactions on Big Data*, 7 (3):535–547, 2019a.

Johnson, J., Douze, M., and Jégou, H. Billion-scale similarity search with GPUs. *IEEE Transactions on Big Data*, 7 (3):535–547, 2019b.

Johnson, L. S., Eddy, S. R., and Portugaly, E. Hidden markov model speed heuristic and iterative hmm search procedure. *BMC bioinformatics*, 11(1):1–8, 2010.

Ju, F., Zhu, J., Shao, B., Kong, L., Liu, T.-Y., Zheng, W.-M., and Bu, D. Copulanet: Learning residue co-evolution directly from multiple sequence alignment for protein structure prediction. *Nature communications*, 12(1):1–9, 2021.

Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., et al. Highly accurate protein structure prediction with alphafold. *Nature*, 596(7873):583–589, 2021.

Kamisetty, H., Ovchinnikov, S., and Baker, D. Assessing the utility of coevolution-based residue-residue contact predictions in a sequence- and structure-rich era. *Proceedings of the National Academy of Sciences of the United States of America*, 2013.

Kaplan, J., McCandlish, S., Henighan, T., Brown, T. B., Chess, B., Child, R., Gray, S., Radford, A., Wu, J., and Amodei, D. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020.

Khandelwal, U., Levy, O., Jurafsky, D., Zettlemoyer, L., and Lewis, M. Generalization through memorization: Nearest neighbor language models. *Learning*, 2019.

Klausen, M. S., Jespersen, M. C., Nielsen, H., Jensen, K. K., Jurtz, V. I., Soenderby, C. K., Sommer, M. O. A., Winther, O., Nielsen, M., Petersen, B., et al. Netsurfp-2.0: Improved prediction of protein structural features by integrated deep learning. *Proteins: Structure, Function, and Bioinformatics*, 87(6):520–527, 2019.

Korendovych, I. V. and DeGrado, W. F. De novo protein design, a retrospective. *Quarterly reviews of biophysics*, 53, 2020.

Lin, Z., Akin, H., Rao, R., Hie, B., Zhu, Z., Lu, W., dos Santos Costa, A., Fazel-Zarandi, M., Sercu, T., Candido, S., et al. Language models of protein sequences at the scale of evolution enable accurate structure prediction. *bioRxiv*, 2022.

Liu, X. Deep recurrent neural network for protein function prediction from sequence. *arXiv preprint arXiv:1701.08318*, 2017.

Marks, D. S., Colwell, L. J., Sheridan, R., Hopf, T. A., Pagnani, A., Zecchina, R., and Sander, C. Protein 3d structure computed from evolutionary sequence variation. *PloS one*, 6(12):e28766, 2011.

Morcos, F., Pagnani, A., Lunt, B., Bertolino, A., Marks, D. S., Sander, C., Zecchina, R., Onuchic, J. N., Hwa, T., and Weigt, M. Direct-coupling analysis of residue coevolution captures native contacts across many protein families. *Proceedings of the National Academy of Sciences*, 108(49):E1293–E1301, 2011.

O'Sullivan, O., Suhre, K., Abergel, C., Higgins, D. G., and Notredame, C. 3dcoffee: combining protein sequences and structures within multiple sequence alignments. *Journal of molecular biology*, 340(2):385–395, 2004.

Pan, X.-Y., Zhang, Y.-N., and Shen, H.-B. Large-scale prediction of human protein- protein interactions from amino acid sequence based on latent topic features. *Journal of proteome research*, 9(10):4992–5001, 2010.

Perdigão, N., Heinrich, J., Stolte, C., Sabir, K. S., Buckley, M. J., Tabor, B., Signal, B., Gloss, B. S., Hammang, C. J., Rost, B., et al. Unexpected features of the dark proteome. *Proceedings of the National Academy of Sciences*, 112 (52):15898–15903, 2015.

Rao, R., Bhattacharya, N., Thomas, N., Duan, Y., Chen, P., Canny, J., Abbeel, P., and Song, Y. Evaluating protein transfer learning with tape. *Advances in neural information processing systems*, 32, 2019.

Rao, R., Meier, J., Sercu, T., Ovchinnikov, S., and Rives, A. Transformer protein language models are unsupervised structure learners. *Biorxiv*, 2020.

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.

Remmert, M., Biegert, A., Hauser, A., and Söding, J. Hhblits: lightning-fast iterative protein sequence searching by hmm-hmm alignment. *Nature methods*, 9(2):173–175, 2012.

Riesselman, A., Shin, J.-E., Kollasch, A., McMahon, C., Simon, E., Sander, C., Manglik, A., Kruse, A., and Marks, D. Accelerating protein design using autoregressive generative models. *BioRxiv*, 757252, 2019.

Rives, A., Goyal, S., Meier, J., Guo, D., Ott, M., Zitnick, C. L., Ma, J., and Fergus, R. Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *Proceedings of the National Academy of Sciences of the United States of America*, 2019.

Rocklin, G. J., Chidyausiku, T. M., Goreshnik, I., Ford, A., Houliston, S., Lemak, A., Carter, L., Ravichandran, R., Mulligan, V. K., Chevalier, A., et al. Global analysis of protein folding using massively parallel design, synthesis, and testing. *Science*, 357(6347):168–175, 2017.

Roy, A., Kucukural, A., and Zhang, Y. I-tasser: a unified platform for automated protein structure and function prediction. *Nature protocols*, 5(4):725–738, 2010.

Sadowski, M. and Jones, D. The sequence–structure relationship and protein function prediction. *Current opinion in structural biology*, 19(3):357–362, 2009.

Stefani, M. Protein misfolding and aggregation: new examples in medicine and biology of the dark side of the protein world. *Biochimica et biophysica acta (BBA)-Molecular basis of disease*, 1739(1):5–25, 2004.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Wang, D., Liu, S., Wang, H., Song, L., Tang, J., Le, S., Grau, B. C., and Liu, Q. Augmenting message passing by retrieving similar graphs. *arXiv preprint arXiv:2206.00362*, 2022.

Wu, R., Ding, F., Wang, R., Shen, R., Zhang, X., Luo, S., Su, C., Wu, Z., Xie, Q., Berger, B., et al. High-resolution de novo structure prediction from primary sequence. *BioRxiv*, pp. 2022–07, 2022.

Xia, X., Zhang, S., Su, Y., and Sun, Z. Micalign: a sequence-to-structure alignment tool integrating multiple sources of information in conditional random fields. *Bioinformatics*, 25(11):1433–1434, 2009.

Xu, M., Zhang, Z., Lu, J., Zhu, Z., Zhang, Y., Ma, C., Liu, R., and Tang, J. Peer: A comprehensive and multi-task benchmark for protein sequence understanding. *arXiv preprint arXiv:2206.02096*, 2022.

Yang, J., Anishchenko, I., Park, H., Peng, Z., Ovchinnikov, S., and Baker, D. Improved protein structure prediction using predicted interresidue orientations. *Proceedings of the National Academy of Sciences*, 117(3):1496–1503, 2020.

Ye, J., McGinnis, S., and Madden, T. L. Blast: improvements for better sequence analysis. *Nucleic acids research*, 34(suppl_2):W6–W9, 2006.

Zhang, H., Ju, F., Zhu, J., He, L., Shao, B., Zheng, N., and Liu, T.-Y. Co-evolution transformer for protein contact prediction. *Advances in Neural Information Processing Systems*, 34:14252–14263, 2021.

Zhang, N., Bi, Z., Liang, X., Cheng, S., Hong, H., Deng, S., Lian, J., Zhang, Q., and Chen, H. Ontoprotein: Protein pretraining with gene ontology embedding. *arXiv preprint arXiv:2201.11147*, 2022.

A. A Brief Recap on Proteins

Proteins are the end products of the decoding process that starts with the information in cellular DNA. As workhorses of the cell, proteins compose structural and motor elements in the cell, and they serve as the catalysts for virtually every biochemical reaction that occurs in living things. This incredible array of functions derives from a startlingly simple code that specifies a hugely diverse set of structures.

In fact, each gene in cellular DNA contains the code for a unique protein structure. Not only are these proteins assembled with different amino acid sequences, but they also are held together by different bonds and folded into a variety of three-dimensional structures. The folded shape, or conformation, depends directly on the linear amino acid sequence of the protein.

1. What are proteins made of?

20 kinds of amino acids. Within a protein, multiple amino acids are linked together by peptide bonds, thereby forming a long chain.

2. Protein structures There are four levels of structures:

- Primary structure: amino acids sequence
- Secondary structure: stable folding patterns, including Alpha Helix, Beta Sheet.
- Tertiary structure: ensemble of formations and folds in a single linear chain of amino acids
- macromolecules with multiple polypeptide chains or subunits

3. Protein Homology Protein homology is defined as shared ancestry in the evolutionary history of life. There exists different kinds of homology, including orthologous homology that may be similar function proteins across species (human and mice α-goblin), and paralogous homology that is the result of mutations (human α-goblin and β-goblin). Homologies result in conservative parts in protein sequences, or leads to similar structures and functions.

4. Multiple Sequence Alignments A method used to determine conservative regions and find homologous sequences. An illustration is given here to show how sequences are aligned.

B. Overview of Previous Protein Representation Augmentation Methods

Below we introduce several state-of-the-art evolution augmentation methods for protein representation learning. These methods rely on MSA as input to extract representations. We use x to denote a target protein and its MSA containing N homologous proteins.

Potts Model (Balakrishnan et al., 2011). This line of research fits a Markov Random Field to the underlying MSA with likelihood maximization. This approach is different from other protein representation learning methods as it only learns a pairwise score for residues contact prediction. We will focus on other methods that augment protein representations that can be used for diverse downstream predictions.

Co-evolution Aggregator (Yang et al., 2020; Ju et al., 2021). One way to build an evolution informed representation is to use a MSA encoder to obtain the co-evolution related statistics. By applying MSA encoder on the n-th homologous protein in the MSA, we can get a total of $L \times d$ embeddings R_n, each position is a d channel one-hot embedding indicating the amino acid type. We use w_n to denote the weight from R_n when computing the token representation h_i:

$$h_i = \frac{1}{M_{eff}} \sum_{n=1}^{N} w_n R_n(i), \tag{5}$$

where $M_{eff} = \sum_{n=1}^{N} w_n$ and $w_n = \frac{1}{N}$. For contact prediction, pair co-evolution representation are computed in a similar way from the hadamard product:

$$h_{ij} = \frac{1}{M_{eff}} \sum_{n=1}^{N} w_n R_n(i) \bigotimes R_n(j). \tag{6}$$

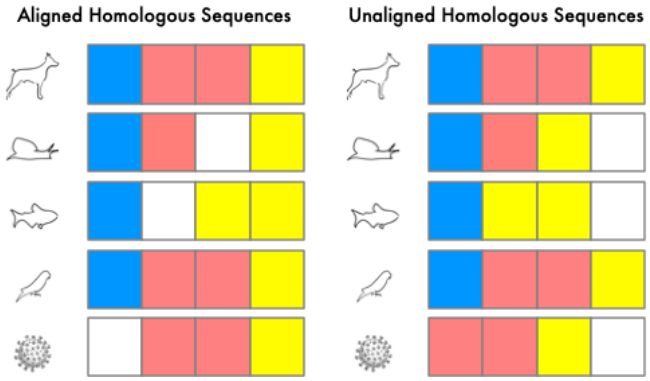

Figure 6. Illustrated difference of aligned and unaligned homologous sequences.

Ensembling Over MSA (Rao et al., 2020). This approach aligns and ensembles representations of homologous sequences. Consider the encoder extract the same token representations for unaligned and aligned sequences. The ensembled token representation is:

$$h_i = \frac{1}{N}\sum_{n=1}^{N} R_n(i), h_{ij} = \frac{1}{N}\sum_{n=1}^{N} \sigma(\frac{R_n(i)W_Q(R_n(j)W_K)^T}{N\sqrt{d}}). \tag{7}$$

MSA Transformer (Rao et al., 2021) In each transformer layer, a tied row attention encoder extracts the dense representation R_n, then a column attention encoder

$$R_s(i) = \sum_{n=1}^{N} \sigma(\frac{R_s(i)W_Q(R_n(i)W_K)^T}{N\sqrt{d}})R_n(i)W_V. \tag{8}$$

C. Experiment Setups

C.1. Introduction to the datasets

Secondary structure prediction (SSP, 8-class) aims to predict the secondary structure of proteins, which indicates the local structures. *Contact prediction* predicts the long-range (distance >6) residue-residue contact, which measures the ability of models to capture global tertiary structures. Homology prediction aims to predict the fold label of any given protein, which indicates the evolutionary relationship of proteins. *Stability* prediction is a protein engineering task, which measures the change in stability w.r.t. residue mutations. *Subcellular Localization (Loc)* prediction predicts the local environment of proteins in the cell, which is closely related to protein functions and roles in biological processes. *Protein protein interaction (PPI)* predicts whether two proteins interact with each other, which is crucial for protein function understanding and drug discovery.

C.2. Retriever and MSA Details

We adopt Faiss (Johnson et al., 2019b) indexing to accelerate the retrieval process by clustering the pre-trained dense vectors. In our implementation, we use the Inverted file with Product Quantizer encoding Indexing and set the size of quantized vectors to 64, the number of centroids to 4096, and the number of probes to 8. During retrieval, L2 distances are used to measure sequence similarity. The index is first trained on .5% of all retrieval data and then add all vectors. For MSA datasets, We use HHblits (Remmert et al., 2012) to perform alignment, and the iteration and E-value thresholds of HHblits are set as 3 and 1.

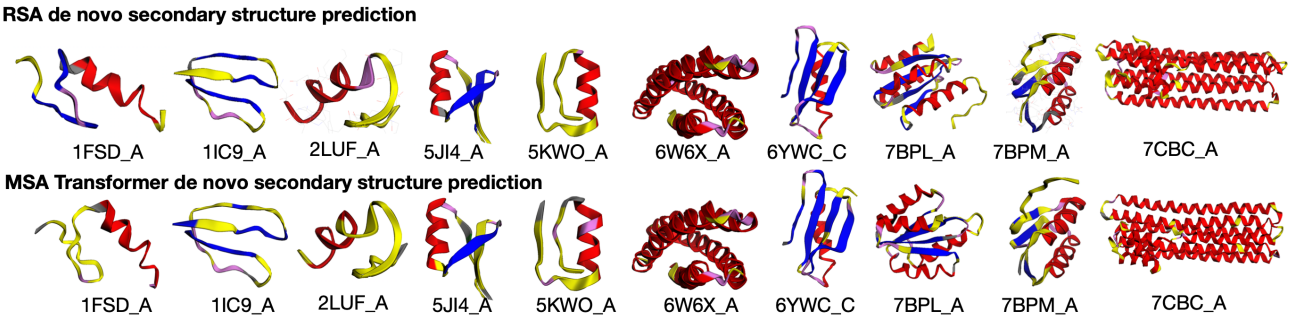

Figure 7. Prediction of Secondary Structure on De Novo Dataset. Each color corresponds to a different secondary structure.

D. Supplementary Experiment Analysis

D.1. Baselines

Protein representation learning benefits from knowledge augmentations. In this part, we examine the performance of three types of baseline models. As shown in Table 3, structure and evolution-related tasks all benefit greatly from pre-training, with over 20% improvement in contact prediction and over 40% improvement in homology prediction. Also, we observe that all kinds of knowledge-augmentation methods improve performance on a few downstream tasks. Though based purely on MSA information, Potts model shows competitive performance to vanilla pre-trained models. MSA Transformer with depth=16 MSA input also sees 12% improvement on its no-MSA input performance. OntoProtein also improves on homology prediction and stability prediction, since knowledge graph enhancement is more suitable to function prediction than structure understanding. PMLM is the SOTA model on both structure and evolution-related tasks through co-evolution pre-training on Pfam database. This trend shows that current scale ($<$1 Billion parameters) pre-trained models still need knowledge augmentations to reach SOTA, and evolutionary knowledge is especially important for downstream prediction.

D.2. Domain Adaptation Analysis

In this section, we perform additional analysis on secondary structure prediction tasks. We perform training on NetSurfP-2.0(Klausen et al., 2019) training set and test on two datasets with domain gaps. On CASP12, RSA marginally outperforms other baselines, as shown in Table 8. We also test on 10 de novo proteins (6YWC, 2LUF, 7BPM, 7BPL, 7CBC, 1FSD, 1IC9, 5JI4, 5KWO, 6W6X). Since we didn't find secondary structure labels for these proteins, we provide visualization in Figure 7 which shows that our model has an obvious overhead over MSA Transformer on predicting geometric components.

Table 8. The domain adaptation performance of models on CASP12 secondary structure prediction.

Method	CASP12
ProtBERT	0.628
MSA Transformer	0.621
Accelerated MSA Transformer	0.620
RSA (ProtBERT backbone)	**0.631**

E. Dataset details

E.1. Downstream tasks

Table 9 gives the details for the datasets.

Table 9. Overview for datasets in downstream tasks

Task Name	Dataset source	#train sequences	#test sequences
Secondary Structure Prediction	NetSurfP-2.0 (Klausen et al., 2019)	8,678	513
Contact Prediction	ProteinNet (AlQuraishi, 2019)	25,299	40
Remote Homology Prediction	Deepsf (Hou et al., 2018)	12,312	718
Stability Prediction	Rocklin's Dataset (Rocklin et al., 2017)	53,571	12,851
Subcellular Localization	DeepLoc (Almagro Armenteros et al., 2017)	8,945	2,768
Protein Protein Interaction	Pan's Dataset (Pan et al., 2010)	6,844	227

E.2. De Novo Protein Dataset

We follow Chowdhury et al. (2022) to curate a de novo dataset of 108 proteins from Protein Data Bank (Bank, 2022). These proteins are originally designed de novo using computationally parametrized energy functions and are well-suited for out-of-domain tests. Note that different from orphan dataset, MSA can be built for this dataset, though showing a decline in quality.

F. Additional Visualization of Retrieved Sequence 3D Structure

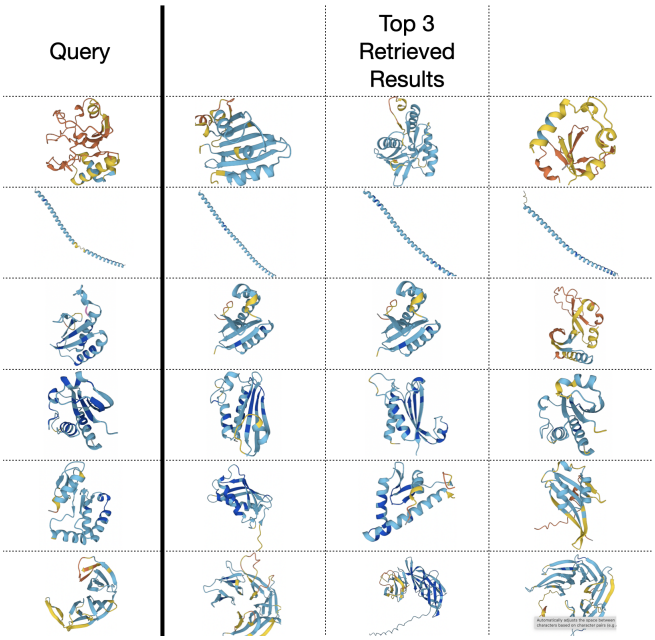

Figure 8. Query and Retrieved Sequence Structures

As shown in Figure 8, we random picked a few more examples to illustrate the structural similarity between query protein and retrieval proteins.