

# RAG-ESM: Improving Pretrained Protein Language Models via Sequence Retrieval

Damiano Sgarbossa<sup>✉\*</sup> and Anne-Florence Bitbol<sup>✉†</sup>

*Institute of Bioengineering, School of Life Sciences, EPFL, CH-1015 Lausanne, Switzerland  
and SIB Swiss Institute of Bioinformatics, CH-1015 Lausanne, Switzerland*



(Received 6 April 2025; accepted 30 July 2025; published 19 August 2025)

Protein language models are significantly advancing the modeling of sequence-function relationships. However, most of them are not directly informed of homology and evolutionary relationships between protein sequences. Here, we propose a method to make them homology aware. We introduce RAG-ESM, a retrieval-augmented-generation (RAG) framework that allows us to condition pretrained ESM2 protein language models on homologous sequences, using a minimal number of additional cross-attention parameters and minimal computational cost. We show that RAG-ESM models outperform larger ESM2 models for masked amino acid prediction. We find that sequence alignment capabilities spontaneously emerge in specific cross-attention heads of RAG-ESM. By using a discrete diffusion objective for training, and by conditioning on homologs during inference, RAG-ESM reaches state-of-the-art performance for conditional protein sequence generation and motif scaffolding, among sequence-based models. Our method thus possesses strong potential for scalable, efficient and controlled protein engineering.

DOI: [10.1103/db1b-hy16](https://doi.org/10.1103/db1b-hy16)

## I. INTRODUCTION

The three-dimensional structures and biological functions of proteins are encoded in their amino-acid sequences. Homologous proteins share a common ancestry, and have similar structure and function. Natural selection for function gives rise to statistical signatures in their sequences. Homology and evolutionary information are thus extremely useful for modeling the sequence-function relationship, and for protein engineering and mutational effect prediction. Experimental approaches like directed evolution and mutational scanning are usually restricted to the local neighborhood of an existing protein sequence. However, the expansion of large-scale databases, such as UniProt [1], has facilitated computational modeling leveraging evolutionary diversity [2–6].

Language models trained on large ensembles of protein sequences produce representations of proteins that correlate with their function [7–11], and enable sequence generation [11,12] and mutational effect prediction [13,14]. These models can be recurrent neural networks [15], transformers [9] or state space models [16], and are trained with objectives such as masked language modeling, autoregressive generation, or discrete diffusion [17,18]. While most of these protein language models (pLMs) were trained on unstructured ensembles of single sequences, some were trained on multiple sequence alignments (MSAs) of homologous proteins, and can thus directly exploit evolutionary diversity and functional con-

straints. MSA-based pLMs include MSA Transformer [19] and AlphaFold2's EvoFormer [20]. Despite being trained only on sequences, MSA Transformer was shown to encode structural contacts between amino acids in its row attention heads [19]. The training of MSA-based pLMs to fill in masked amino acids using the surrounding context of an MSA allows them to efficiently capture coevolution between amino acids due to structural constraints, with far fewer parameters than models based on single sequences [19,21]. These models are an important ingredient of AlphaFold's major advance in structure prediction [20,22]. However, MSA-based models are memory-intensive [19], and may inherit the imperfections of MSAs [23]. Recently, autoregressive models have been trained on concatenations of nonaligned homologs [16,24]. However, starting from long sequences of concatenated homologs poses memory issues for transformer-based models.

We posit that augmenting pretrained single-sequence pLMs by using homology could improve their performance, thereby combining the advantages of single-sequence and MSA-based models. To investigate this, we use Retrieval-Augmented Generation (RAG) to improve the pLM ESM2 [21], by using homologous sequences as the retrieved external sources. RAG allows us to improve the accuracy of generative large language models by integrating relevant data from external knowledge sources (e.g., domain-specific information) into the generation process [25,26]. Our model RAG-ESM outperforms substantially larger ESM2 models at masked amino acid prediction. Furthermore, we find that sequence alignment capabilities emerge in cross-attention heads. By using a discrete diffusion objective during training and conditioning on homologs during inference, we perform conditional protein sequence generation and motif scaffolding using RAG-ESM, and obtain state-of-the-art performance. We further show that retrieval yields significant improvements on the structural and evolutionary fidelity of generated sequences. RAG-ESM builds on pretrained pLMs, and improves

\*Contact author: [damiano.sgarbossa@epfl.ch](mailto:damiano.sgarbossa@epfl.ch)

†Contact author: [anne-florence.bitbol@epfl.ch](mailto:anne-florence.bitbol@epfl.ch)

*Published by the American Physical Society under the terms of the Creative Commons Attribution 4.0 International license. Further distribution of this work must maintain attribution to the author(s) and the published article's title, journal citation, and DOI.*

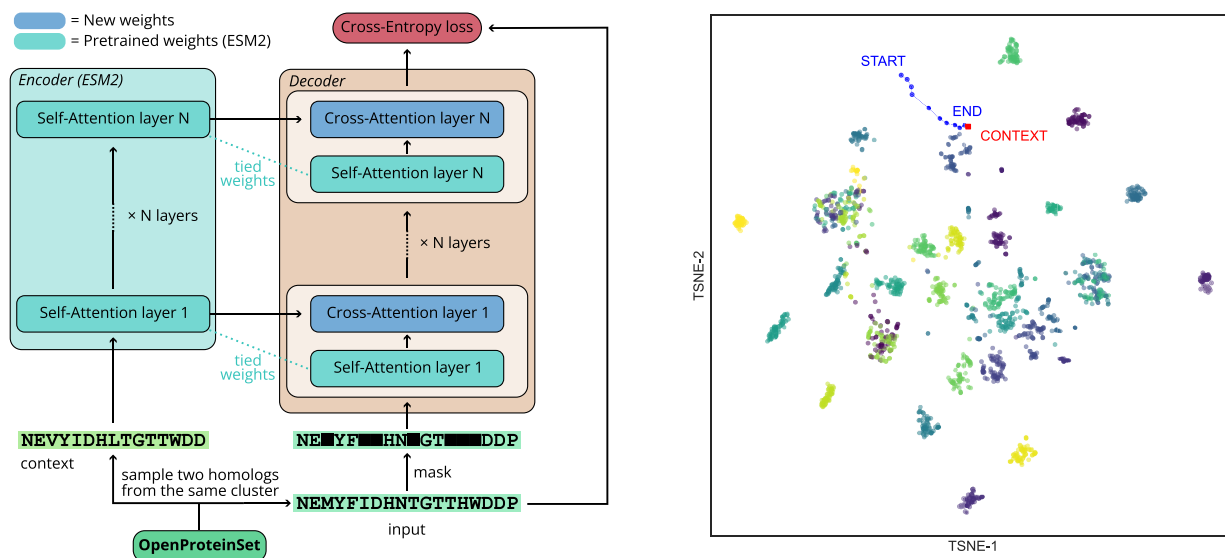


FIG. 1. Left: Schematic of the RAG-ESM architecture. “Self-Attention layer” and “Cross-Attention layer” denote the concatenation of attention and feed-forward layers. Right: Two-dimensional t-SNE visualization of the ESM2 (650M) embeddings of protein sequences from different clusters in the test set (different colors from yellow to purple). Blue: path of a sequence generated from scratch using RAG-ESM (165M) by denoising (see Supplemental Material Sec. C [29]); START: initial sequence of only <mask> tokens; END: sequence obtained after denoising; CONTEXT: sequence given as context.

their performance and efficiency. It thus possesses strong potential for scalable and efficient protein engineering.

## II. METHODS

### A. Summary of our contributions

In this work, we demonstrate that retrieval techniques can improve pretrained single-sequence pLMs, such as ESM2 [21]. Inspired by Ref. [27], where pretrained pLMs were conditioned on structural information through a structure encoder and cross attention on the final layer, we similarly condition pretrained pLMs, but using sequence information [see Fig. 1 (left) and Sec. II B].

We introduce RAG-ESM, an encoder-decoder model trained with masked language modeling on pairs of homologous protein sequences. The model takes as input both a masked sequence and a context sequence. The latter can be one of the homologs of the former, or any sequence with specific desired properties (e.g., belonging to a particular family, having a specific function, or binding to a specific domain). The context sequence is embedded using an encoder model. The resulting embeddings are provided to a decoder, in order to improve its predictions of the masked amino acids in the input sequence. We show that

(1) By conditioning on a homologous sequence, we drastically improve the performance of pretrained pLMs at predicting masked amino acids. The computational costs are minimal (50 to 120 GPU hours, depending on model size) and the number of additional parameters is small, opening new directions for more efficient and scalable future models.

(2) By training the models using a discrete diffusion objective (see Sec. II B) and conditioning, during inference, on sequences with specific desired properties, we transform a

pretrained single-sequence masked language model into one that can perform conditional generation.

This second contribution addresses a key limitation in the generative abilities of most single-sequence pLMs: their lack of control over generated outputs. Conditioning on sequence provides an alternative to conditioning on control tags, e.g., gene ontology terms [11,29], and to multimodal pLMs using structural and functional features [18,30]. By providing a representation of a context sequence, we guide the model to a sample from a specific region of sequence space, significantly reducing the dimensionality of the search space. Figure 1 (right) shows that during the denoising process (described in Supplemental Material Sec. C [28]), the model starts from a position of the embedding space corresponding to the fully masked sequence, and quickly converges to a position close to the context sequence embedding and its neighboring homologs.

Our approach reduces the need for the model to allocate parameters for memorizing protein family information [8,30,31]. Instead, this information can be recovered from the context embeddings. We find that this enables performance comparable to much larger models, with substantially fewer parameters.

Furthermore, to improve generation quality, we introduce an Error Correction strategy that allows the model to revise previously sampled residues during denoising. This leverages a masked language modeling (MLM)-like training objective (see Sec. II B and Supplemental Material Sec. C [28]), which exposes the model to corrupted inputs and trains it to detect and correct unlikely residues, even when they are not masked. In Table S6, we compare different sequence generation strategies used during denoising, showing that our proposed Error Correction method improves the quality of the output sequences.

## B. Architecture and training

The RAG-ESM model builds upon a pretrained ESM2 model modified by adding a few additional layers. The architecture is shown in Fig. 1 (left). The model comprises two main modules:

(1) An encoder module, which corresponds to the pretrained ESM2 model, and computes the embeddings of the (unmasked) context sequences.

(2) A decoder module, built starting from the pretrained layers of the ESM2 model, adding newly initialized cross-attention layers to some of them. These layers integrate information from the context embeddings (from the encoder) and the input ones (from the decoder) into a single representation. This module takes as input the masked sequence and the embeddings of the context sequence. It provides as outputs the logits for the masked amino acids.

The weights of the ESM2 layers (including both attention and feed-forward layers) used by both encoder and decoder modules are tied, i.e., the parameters are shared between the two modules. Thus, starting with an  $N$ -parameter ESM2 model, and using  $M$  parameters in the cross-attention layers, our RAG-ESM model has  $N + M$  parameters. In practice,  $M \ll N$ , as we apply cross attention to few layers. Indeed, few layers of cross attention suffice to transfer information between the two sequences' embeddings, see Supplemental Material Sec. B. Thus motivated, we train RAG-ESM models with 12M and 165M parameters, respectively based on the 8M and 150M ESM2 models.

We train our models using the standard cross-entropy loss with a discrete diffusion objective [32]. The masking fraction is sampled from a uniform distribution over  $(0, 1)$  to improve its generation capabilities during the denoising steps (see Supplemental Material Sec. A [29]). We simultaneously fine-tune the pretrained self-attention weights (from ESM2) and train the newly initialized cross-attention weights with different learning rates. At each training instance, a sequence is selected as input, and its closest neighbor (by Hamming distance) in the same cluster is used as context. While the final performance of the model at inference is robust to the similarity level of the context sequences used for training, using the closest neighbor during training accelerates loss convergence, see Supplemental Material Sec. B [29].

All models were trained on pairs of homologs sampled from Uniclust30 clusters [33] present in the filtered version of OpenProteinSet [34], which comprises maximally diverse representative clusters. This dataset includes a total of  $\sim 200$  million protein sequences from UniprotKB [1]. More details are given in Supplemental Material Sec. A [29].

## III. RESULTS

### A. Attending to close homologs drastically improves the perplexity of pLMs

Does retrieval impact the performance of pLMs? How much does this depend on the similarity between context and input sequences? We address these questions by investigating the models' perplexity.

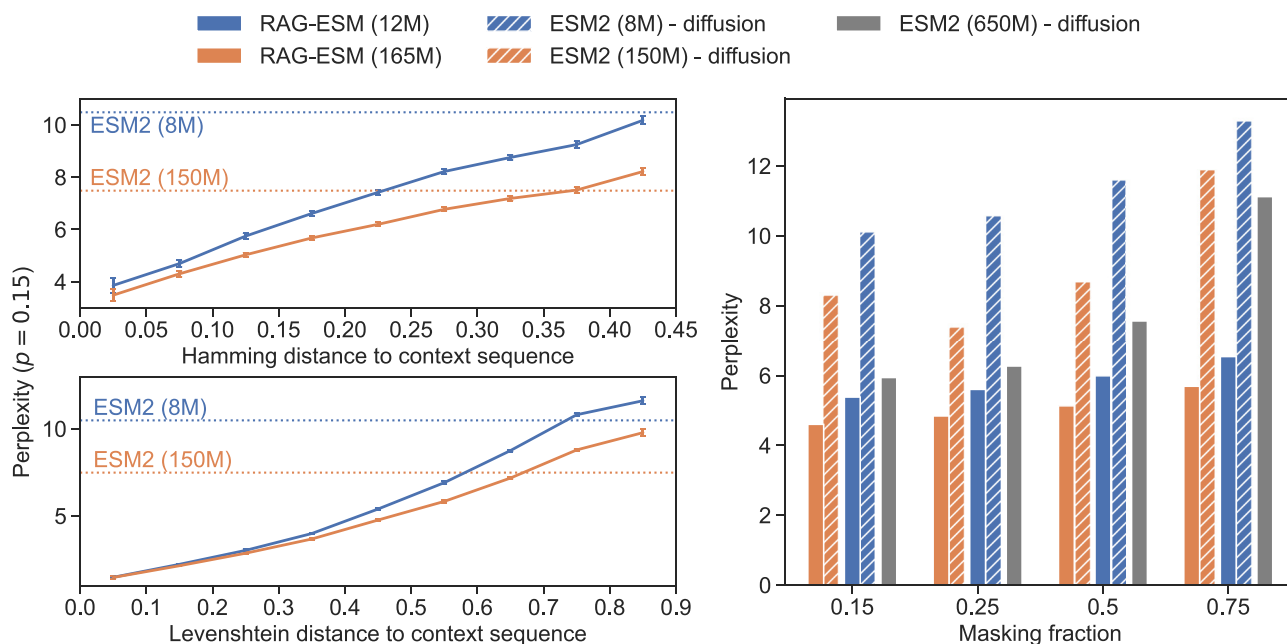


FIG. 2. Left: Perplexity of the RAG-ESM models on the standard MLM task ( $p = 0.15$ ) versus distance between input and context sequences (both from the same cluster). We measure distance using both Hamming and Levenshtein distances (resp. between sequences aligned using the pairwise Smith-Waterman algorithm, and between unaligned sequences), binned in intervals of 0.1. Right: Perplexity of ESM2 models and RAG-ESM models, both trained on the discrete diffusion task, for different masking fractions, when the closest homolog is provided as context to RAG-ESM (mean Hamming dist  $\sim 0.36$ ). Perplexity is measured on 10 000 sequences sampled from the validation set. ESM2 (650M), fine-tuned on the discrete diffusion objective, serves as baseline (see Table S5 for a comparison with RAG-ESM models trained on other objectives).

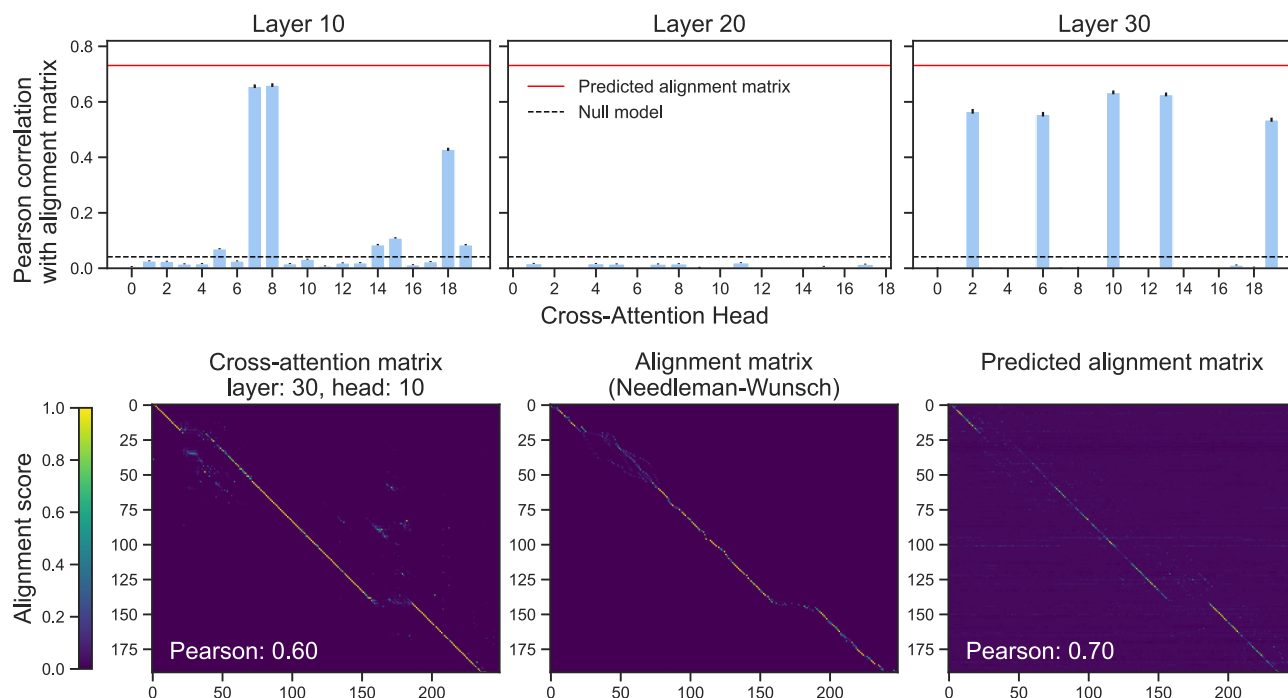


FIG. 3. Top: Pearson correlation between alignment matrices obtained by aligning input and context sequences (using the Needleman-Wunsch algorithm) and each of the 60 cross-attention heads from RAG-ESM (165M) – 20 per layer across layers 10, 20, and 30. Blue and black: mean and standard error of the correlations for each head. Red line: mean Pearson correlation obtained for the logistic regression trained on 100 examples. Dashed black line: mean Pearson correlation between the alignment matrices and the same matrices with shuffled entries (null model). Results are obtained by feeding one pair of homologs from each of the 1000 clusters in the test set to the RAG-ESM model, without masking the input sequences. All correlations are computed over the top  $2L$  matrix entries, where  $L$  is sequence length. Bottom: Visual comparison of the Needleman-Wunsch alignment matrix (center) with one of the cross-attention heads of RAG-ESM (left) and with the alignment matrix predicted by the logistic regression (right), for a random pair of homologous sequences sampled from a cluster in the test set. Their Hamming distance is 0.73.

Figure 2 (left) shows the effect of the distance between context and input sequences on model performance. When the two sequences are the same, the perplexity of the model is 1, i.e., the model perfectly predicts the masked amino acids by looking up those in the context sequence. Perplexity increases with the distance between the two sequences. RAG-ESM reaches perplexities similar and sometimes slightly larger than the base ESM2 model when very different sequences from the same cluster are used as context and input.

In Fig. 2 (right), we compare RAG-ESM models with the base ESM2 models they are built on. We find that the usage of homology information leads to large performance improvements. We compare the pretrained ESM2 (8M) and ESM2 (150M), fine-tuned on the discrete diffusion task, to their RAG-ESM counterparts with 12M and 165M parameters. We obtain, respectively, a 48% and 43% decrease in perplexity [averaged over the different masking fractions considered in Fig. 2 (right)] with respect to the base models, when using the closest homolog as context. Performance decreases when using more distant homologs (see Table S5), but it remains at least as good as the one of ESM2 (650M). Furthermore, even the smaller RAG-ESM (12M) model reaches a lower perplexity than the much larger ESM2 (650M) model. This shows that using homologs information can help to substantially decrease the size of models, thus fostering computational efficiency.

## B. Alignment capabilities naturally emerge in the cross-attention heads

How interpretable are cross-attention heads in our model? Do they learn to align input and context sequences without explicit alignment information during training? To address these questions, we examine the cross-attention coefficients calculated by RAG-ESM when feeding it pairs of homologs, without masking input sequences. Figure 3 (top) shows that some cross-attention heads of RAG-ESM (165M) exhibit a high Pearson correlation  $\rho$  with pairwise sequence alignment matrices computed using a differentiable formulation proposed by Ref. [35], which was shown to converge to the Needleman-Wunsch alignment. Specifically, seven of these heads feature  $\rho > 0.5$ . Note that given the sparsity of these matrices, we compute correlations  $\rho$  over their top  $2L$  entries, where  $L$  is sequence length. Notably, the alignment-specialized heads appear only in the first and last cross-attention layers, while the middle layer shows little to no significant correlation. Figure 3 (bottom) shows an alignment matrix and cross-attention matrices for two homologous sequences.

To further assess the ability of attention heads at capturing alignment matrices, we trained a logistic regression model (with predictions ranging continuously between 0 and 1) on the 60 cross-attention heads to predict the alignment matrices. Using 100 samples from the test set for training, and evaluating on the remaining 900 samples, we obtained an average

TABLE I. Quality scores for generated and natural sequences from 1000 different clusters in the test set. We report the median and standard deviation of each distribution. Both generated and natural sequences are filtered by computing the ESM2 (650M) pseudoperplexity and retaining only those with perplexity  $< 8$  (approximately 650 sequences out of 1000, both for natural and generated sequences). HMMER scores are normalized by dividing each of them by the score of the corresponding context sequence. For natural sequences, Hamming distances to the context are computed for randomly sampled sequences, either from the same cluster as the context (“Homolog”) or from a random different cluster (“Random”).  $\uparrow$  (resp.  $\downarrow$ ) indicates that higher (resp. lower) scores are better.

| Score                                   | Generated       |                  | Natural         |                 |
|-----------------------------------------|-----------------|------------------|-----------------|-----------------|
|                                         | RAG-ESM (165M)  | ESM2 (650M)      | Homolog         | Random          |
| ESMFold pLDDT ( $\uparrow$ )            | $0.76 \pm 0.12$ | $0.52 \pm 0.16$  | $0.81 \pm 0.15$ |                 |
| scPerplexity ( $\downarrow$ )           | $2.58 \pm 0.44$ | $2.96 \pm 0.68$  | $2.52 \pm 0.43$ |                 |
| ESM2 (650M) perplexity ( $\downarrow$ ) | $6.69 \pm 2.66$ | $7.16 \pm 2.47$  | $6.06 \pm 3.03$ |                 |
| RMSD with context ( $\downarrow$ )      | $4.49 \pm 5.41$ | $13.13 \pm 7.76$ | $9.68 \pm 6.0$  | $11.85 \pm 4.5$ |
| TMScore with context ( $\uparrow$ )     | $0.49 \pm 0.26$ | $0.14 \pm 0.05$  | $0.19 \pm 0.15$ | $0.15 \pm 0.05$ |
| HMMER ( $\uparrow$ )                    | $0.96 \pm 0.76$ | $-0.01 \pm 0.25$ | $0.92 \pm 0.62$ | $-0.01 \pm 0.1$ |
| Hamming to context ( $-$ )              | $0.50 \pm 0.21$ | $0.74 \pm 0.04$  | $0.50 \pm 0.16$ | $0.77 \pm 0.05$ |

Pearson correlation of  $\rho = 0.73$  between the alignment matrices predicted by the logistic model and the original ones, see Fig. 3 (top). This result is reminiscent of [19], where logistic regression on row attention matrices was shown to capture contact maps, and of Ref. [36], where logistic regression on column attention matrices was found to predict Hamming distances between sequences.

These results suggest that the cross attention between the input and context sequences enables the model to implicitly perform sequence alignment, despite not being trained on explicitly aligned sequences. In other words, the training objective of RAG-ESM encourages the extraction of informative signals from homologous context sequences. This leads to the emergence of specialized cross-attention heads that tend to effectively align the two sequences.

### C. Diffusion-based denoising enables conditional protein sequence generation

RAG-ESM can generate protein sequences conditioned on homologous context via a diffusion-based denoising process that progressively reveals masked parts of the input sequences (see Supplemental Material Sec. C for details [29]). How does retrieval influence generative performance? To answer this, we compare sequences generated by RAG-ESM (165M) with those produced by a diffusion fine-tuned ESM2 (650M) baseline, which does not directly use homolog information, and with natural sequences from the test set. The natural sequences consist of (i) homologous ones sampled from the same clusters as the generated sequences, and (ii) nonhomologous ones sampled from random clusters.

To assess the quality of the generated sequences, we employ several established metrics. Sequence confidence is measured via the pseudo-perplexity of the nonfine-tuned ESM2 (650M), while structural confidence is quantified by the ESMFold pLDDT [21] and the ProteinMPNN self-consistency perplexity (scPerplexity, computed from ESMFold-predicted structures, following Ref. [17]). In addition, structural similarity between the predicted structures of generated and context sequences is evaluated using RMSD (Root Mean Square Deviation) and TMScore (after structural alignment) [37], and homology is assessed using a Hidden Markov Model (HMM) trained on the cluster’s MSA, via HMMER scores [38]. Finally, we use the Hamming distance to quantify how much each generated sequence diverges from the corresponding context sequence.

The results in Table I indicate that RAG-ESM (165M) significantly outperforms the larger ESM2 baseline for conditional sequence generation. This finding confirms the benefit of incorporating homologs to improve model performance during inference. Notably, RAG-ESM-generated sequences have scores closer to those of natural sequences, and sometimes exceed them. Specifically, the HMMER scores of RAG-ESM-generated sequences closely approximate those of natural homologs, highlighting the model’s ability to preserve key characteristics of protein families. Moreover, the Hamming distances between generated sequences and their corresponding context sequences are comparable to those observed between natural homologs, indicating that our method produces sequences that appropriately differ from the provided context. In terms of structural plausibility, the sequences generated by RAG-ESM feature ESMFold pLDDT

TABLE II. Performance of RAG-ESM and other models for homolog-conditioned generation. We report the median and standard deviation of each distribution. Two structural scores are presented: the pLDDT from ESMFold [21] and the scPerplexity from ProteinMPNN [39], evaluated on a set of 250 protein sequences generated using each of the models, each from a distinct cluster in the test set. For other models than RAG-ESM, results were retrieved from the Zenodo archive associated with the EvoDiff paper [17] and the ProtMamba paper [16].  $\uparrow$  (resp.  $\downarrow$ ) indicates that higher (resp. lower) scores are better.

| Model                         | RAG-ESM                           | ProtMamba       | EvoDiff-MSA     | MSA Trans.      | Potts           | Natural         |
|-------------------------------|-----------------------------------|-----------------|-----------------|-----------------|-----------------|-----------------|
| pLDDT ( $\uparrow$ )          | <b><math>0.76 \pm 0.12</math></b> | $0.75 \pm 0.13$ | $0.60 \pm 0.16$ | $0.54 \pm 0.18$ | $0.56 \pm 0.14$ | $0.81 \pm 0.15$ |
| scPerplexity ( $\downarrow$ ) | <b><math>2.58 \pm 0.44</math></b> | $2.63 \pm 0.45$ | $3.17 \pm 0.58$ | $3.37 \pm 0.64$ | $3.17 \pm 0.51$ | $2.52 \pm 0.43$ |

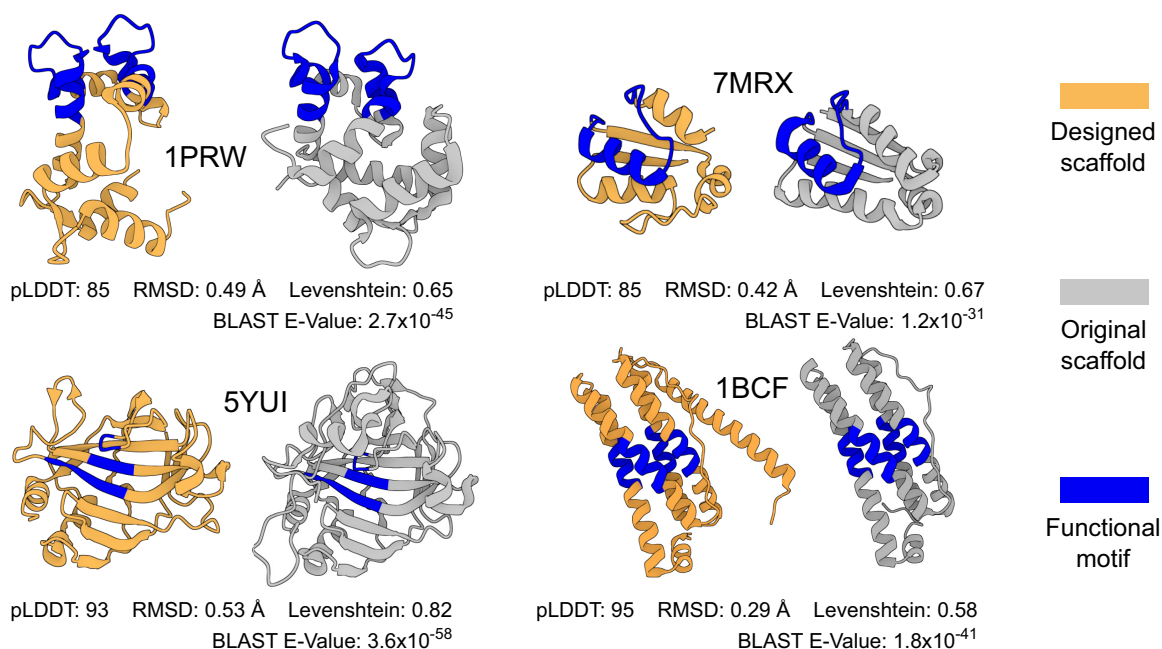


FIG. 4. Examples of successfully scaffolded motifs. The structure of each motif is highlighted in blue, while gray denotes the reference structures from the PDB, and orange denotes the ESMFold-predicted structures for the designed scaffolds. For each design, we report the ESMFold pLDDT (higher is better), the RMSD—restricted to the motif region—between generated and reference structure (lower is better), the Levenshtein distance between generated and reference sequence (higher is better), and the BLAST E-Value score when searching the UniProtKB database [1]. Note that the similarity between the designed sequences and the closest natural sequence from UniprotKB, identified by the BLAST search, are the following: 1BCF: 59.5%; 7MRX: 68.4%; 5YUI: 56.7%; 1PRW: 62.8%.

scores and ProteinMPNN self-consistency perplexities that are similar to those of natural sequences, and superior to those generated by the larger ESM2 baseline. Although the ESM2 pseudoperplexity for RAG-ESM is marginally less favorable than that of natural sequences, it still outperforms the ESM2 baseline. Additionally, RAG-ESM produces sequences that are structurally more similar to their context than ESM2, as evidenced by significantly lower RMSD and higher TMScore values compared to both natural homologs and ESM2-generated sequences. Thus, the conditional generation strategy yields enhanced structural performance.

Table II further compares the performance of RAG-ESM for conditional generation with other protein language models. ProtMamba [16], currently state-of-the-art for conditional generation, is a Mamba-based autoregressive model that leverages evolutionary information during inference by concatenating multiple unaligned homologous sequences as context. EvoDiff-MSA is an MSA-based diffusion protein language model whose capabilities have been experimentally validated [17]. MSA Transformer is the pretrained model used as the starting point for EvoDiff-MSA and was trained with a masked language modeling objective [19]. Finally, we consider Potts models trained on the MSAs of specific families, an experimentally validated generative model inspired by statistical physics [5]. We compare RAG-ESM with these baseline models using the performance reported in Refs. [17] and [16]. Strikingly, RAG-ESM achieves the highest median ESMFold pLDDT and the lowest ProteinMPNN scPerplexity among the generative models considered. Thus, RAG-ESM performs as well as the state-of-the-art method ProtMamba,

even if it conditions on single sequences instead of larger sets of homologs from the same cluster. Furthermore, RAG-ESM exceeds the performance of other existing models for homolog-conditioned generation.

Together, these results, along with the high cosine similarity observed between ESM2 (650M) embeddings of generated and context sequences (Fig. S1), demonstrate that the diffusion-based denoising strategy of RAG-ESM effectively leverages homologous information for conditional generation. The approach yields protein sequences that are both evolutionarily and structurally plausible, performing similarly to natural sequences across the evaluated metrics. This positions RAG-ESM as a promising alternative to existing generation methods, particularly in applications where preserving context-specific structural and evolutionary characteristics is critical.

#### D. RAG-ESM reaches state-of-the-art performance for motif scaffolding

Motivated by the promising results in conditional sequence generation, we assess the generative performance of RAG-ESM (165M) for the motif-scaffolding task introduced by Ref. [40]. The goal is to design protein scaffolds that accommodate a fixed functional motif, such as a binding or function-determining region, by generating the surrounding sequence that supports its structural integrity. Figure 4 shows examples of motifs that were successfully scaffolded via denoising with RAG-ESM.

TABLE III. Performance comparison of different models on the motif-scaffolding task. For each of the 17 motifs considered in [40], labeled by the associated PDB entry, we report the success rate (out of 100 attempts for each target) of scaffold generation for five models: RAG-ESM (165M), DPLM (650M), EvoDiff-MSA, RFDiffusion, and ESM3 (1.4B). Scaffold sequences are generated with lengths sampled uniformly from [50, 100] plus the motif length, and subsequently folded using ESMFold, similarly to Ref. [17]. A design is considered successful if its folded structure exhibits a pLDDT score greater than 0.70 and an RMSD to the motif (in the reference structure) smaller than 1 Angstrom. We also report the overall pass rate (i.e., the number of tasks with at least one successful design), as well as the mean and median success rates. Both are reported because the distribution of success rates across motifs is highly skewed. Models with the best performance are in bold, and those with the second-best one are underlined. For ESM3, we show values reported in Ref. [42], thus omitting 6vw1.

| PDB ID            | Sequence-based |             | Alignment-based<br>EvoDiff-MSA | Structure-based<br>RFDiffusion | Multimodal<br>ESM3 |
|-------------------|----------------|-------------|--------------------------------|--------------------------------|--------------------|
|                   | RAG-ESM        | DPLM        |                                |                                |                    |
| 1bcf              | <u>0.99</u>    | <b>1.00</b> | 0.84                           | 1.00                           | 0.89               |
| 1prw              | <u>0.43</u>    | <b>0.75</b> | 0.07                           | 0.08                           | 0.96               |
| 1qjg              | <u>0.01</u>    | 0.00        | <b>0.11</b>                    | 0.00                           | 0.02               |
| 1ycr              | <u>0.05</u>    | <b>0.27</b> | 0.00                           | 0.74                           | 0.41               |
| 2kl8              | <u>0.01</u>    | <u>0.01</u> | <b>0.03</b>                    | 0.88                           | 0.11               |
| 3ixt              | <u>0.11</u>    | <b>0.15</b> | 0.00                           | 0.25                           | 0.18               |
| 4jhw              | <u>0.00</u>    | 0.00        | 0.00                           | 0.00                           | 0.00               |
| 4zyp              | 0.00           | 0.00        | 0.00                           | 0.40                           | 0.03               |
| 5ius              | <b>0.25</b>    | 0.00        | 0.00                           | 0.02                           | 0.00               |
| 5tpn              | 0.00           | 0.00        | 0.00                           | 0.61                           | 0.03               |
| 5trv              | <b>0.13</b>    | 0.00        | <u>0.06</u>                    | 0.22                           | 0.16               |
| 5wn9              | 0.00           | 0.00        | 0.00                           | 0.00                           | 0.02               |
| 5yui              | <u>0.35</u>    | <b>0.94</b> | 0.33                           | 0.00                           | 0.00               |
| 6e6r              | 0.36           | <b>0.79</b> | <u>0.73</u>                    | 0.71                           | 0.24               |
| 6exz              | <u>0.31</u>    | 0.01        | <b>0.81</b>                    | 0.42                           | 0.31               |
| 6vw1              | <u>0.01</u>    | 0.00        | <b>0.04</b>                    | 0.69                           | —                  |
| 7mrx              | 0.38           | <u>0.54</u> | <b>0.57</b>                    | 0.07                           | 0.65               |
| Pass rate         | <b>13/17</b>   | 9/17        | <u>10/17</u>                   | 13/17                          | 13/16              |
| Mean succ. rate   | 0.20           | <b>0.26</b> | <u>0.21</u>                    | 0.36                           | 0.25               |
| Median succ. rate | <b>0.11</b>    | 0.01        | <u>0.04</u>                    | 0.25                           | 0.14               |

Here, we follow the procedure proposed in Ref. [17], with one modification: we use ESMFold [21] as the structure prediction model instead of OmegaFold [41], motivated by the higher accuracy of the former. We sample the new scaffolds by conditioning RAG-ESM on the original sequence of each motif and using a temperature of  $T = 0.7$ .

How does RAG-ESM compare to other models for motif scaffolding? Previous approaches include sequence-based methods like DPLM (650M) [18], the alignment-based EvoDiff-MSA [17], structure-based models such as RFDiffusion [40], and large multimodal models incorporating structural information like ESM3 (1.4B) [30]. In Table III, we compare the success rates of RAG-ESM (165M) to these methods for the motif-scaffolding task. Our results show that RAG-ESM outperforms the larger DPLM (650M) model, as well as the MSA-based EvoDiff-MSA. Although the structure-based model RFDiffusion and the multimodal model ESM3 remain superior on some motifs, they are outperformed by RAG-ESM on others. Thus, we envision that RAG-ESM could complement structure-based approaches.

How robust are the comparisons between models presented here? One challenge for detailed comparison between sequence-based and structure-based models is that structure-based models like RFDiffusion follow a different pipeline for motif scaffolding. They usually generate a structural scaffold that is subsequently inverse-folded using ProteinMPNN. We nevertheless included these comparisons here for complete-

ness. Besides, there appear to be some differences in the reported performance of DPLM for motif-scaffolding tasks between [18] and the more recent study [42]. For completeness, in Table S7, we compare our results with the updated DPLM performance [42]. That comparison is even more favorable to RAG-ESM than Table III. Finally, in Table S8, we present a similar comparison as in Table III, but using OmegaFold instead of ESMFold as the structure prediction method. This yields higher success rates due to OmegaFold's higher pLDDT scores, but does not modify our conclusions on the comparisons between models. Thus, our conclusions are robust, and show that RAG-ESM is very promising for motif scaffolding.

#### IV. DISCUSSION

In this work, we introduced RAG-ESM, a sequence retrieval-augmented framework that builds upon a pretrained protein language model trained on an unstructured ensemble of protein sequences, specifically ESM2. Specifically, we augmented ESM2 with a small number of cross-attention layers. RAG-ESM takes as input a masked sequence and as context a homologous sequence, and computes cross-attention between them. We showed that such conditioning on homologous sequences dramatically reduces perplexity, and enables relatively small models to perform on par with much larger ones. We also found that some cross-attention heads

of RAG-ESM spontaneously learn to align input and context sequences. Thus, RAG-ESM efficiently learns sequence homology.

By integrating a discrete diffusion objective, we enabled RAG-ESM to perform conditional protein sequence generation. We showed that RAG-ESM not only achieves state-of-the-art performance in generating novel sequences, but also excels in the motif-scaffolding task among sequence-based models. In fact, it outperforms models that either use significantly more parameters [18] or rely on additional information, such as MSA-based models [17]. An asset of RAG-ESM is that it strongly benefits from using just one homologous sequence as context, while other models that perform homology-conditioned generation require MSAs [19] or large collections of homologs [16,24], and models that condition on control tags need annotated data for training [10,11,43]. This makes RAG-ESM more flexible, enhancing its applicability, e.g., to small protein families and poorly annotated data.

Our work opens several promising future directions. For instance, incorporating additional sources of biological information, such as structural data or functional annotations,

could further improve the controllability and accuracy of generated sequences [30,44–46]. Moreover, exploring more advanced denoising algorithms might enhance the generative capabilities of the model even further [18,32]. Besides, the emergent alignment ability of RAG-ESM suggests its potential for developing new methods for unsupervised protein alignment and evolutionary analysis. Finally, RAG-ESM establishes a scalable, efficient, and versatile protein sequence generation method conditioned on homologous sequences. It thus opens new avenues for controlled protein engineering.

## ACKNOWLEDGMENT

This research was partly funded by the European Research Council (ERC) under the European Union’s Horizon 2020 research and innovation programme (Grant Agreement No. 851173, to A.-F.B.).

## DATA AVAILABILITY

The data that support the findings of this article are openly available [47].

- [1] The UniProt Consortium, UniProt: The universal protein knowledgebase in 2021, *Nucl. Acids Res.* **49**, D480 (2021).
- [2] M. Weigt, R. A. White, H. Szurmant, J. A. Hoch, and T. Hwa, Identification of direct residue contacts in protein-protein interaction by message passing, *Proc. Natl. Acad. Sci. USA* **106**, 67 (2009).
- [3] F. Morcos, A. Pagnani, B. Lunt, A. Bertolino, D. S. Marks, C. Sander, R. Zecchina, J. N. Onuchic, T. Hwa, and M. Weigt, Direct-coupling analysis of residue coevolution captures native contacts across many protein families, *Proc. Natl. Acad. Sci. USA* **108**, E1293 (2011).
- [4] D. S. Marks, L. J. Colwell, R. Sheridan, T. A. Hopf, A. Pagnani, R. Zecchina, and C. Sander, Protein 3D structure computed from evolutionary sequence variation, *PLoS ONE* **6**, e28766 (2011).
- [5] W. P. Russ, M. Figliuzzi, C. Stocker, P. Barrat-Charlaix, M. Socolich, P. Kast, D. Hilvert, R. Monasson, S. Cocco, M. Weigt, and R. Ranganathan, An evolution-based model for designing chorismate mutase enzymes, *Science* **369**, 440 (2020).
- [6] A. Hawkins-Hooker, F. Depardieu, S. Baur, G. Couairon, A. Chen, and D. Bikard, Generating functional protein variants with variational autoencoders, *PLoS Comput. Biol.* **17**, e1008736 (2021).
- [7] A. Elnaggar, M. Heinzinger, C. Dallago, G. Rehawi, Y. Wang, L. Jones, T. Gibbs, T. Feher, C. Angerer, M. Steinegger, D. Bhowmik, and B. Rost, ProtTrans: Towards cracking the language of life’s code through self-supervised deep learning and high performance computing, in *Proceedings of the IEEE Transactions on Pattern Analysis and Machine Intelligence* (IEEE, New York, 2021), p. 1.
- [8] J. Vig, A. Madani, L. R. Varshney, C. Xiong, R. Socher, and N. Rajani, BERTology meets biology: Interpreting attention in protein language models, in *International Conference on Learning Representations* (ICLR, 2021).
- [9] A. Rives, J. Meier, T. Sercu, S. Goyal, Z. Lin, J. Liu, D. Guo, M. Ott, C. L. Zitnick, J. Ma, and R. Fergus, Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences, *Proc. Natl. Acad. Sci. USA* **118**, e2016239118 (2021).
- [10] A. Madani, B. McCann, N. Naik, N. S. Keskar, N. Anand, R. R. Eguchi, P.-S. Huang, and R. Socher, ProGen: Language modeling for protein generation, [bioRxiv: 2020.03.07.982272](https://arxiv.org/abs/2020.03.07.982272).
- [11] A. Madani, B. Krause, E. R. Greene, S. Subramanian, B. P. Mohr, J. M. Holton, J. L. Olmos, C. Xiong, Z. Z. Sun, R. Socher, J. S. Fraser, and N. Naik, Large language models generate functional protein sequences across diverse families, *Nat. Biotechnol.* **41**, 1099 (2023).
- [12] N. Ferruz, S. Schmidt, and B. Höcker, ProtGPT2 is a deep unsupervised language model for protein design, *Nat. Commun.* **13**, 4348 (2022).
- [13] J. Meier, R. Rao, R. Verkuil, J. Liu, T. Sercu, and A. Rives, Language models enable zero-shot prediction of the effects of mutations on protein function, *Adv. Neural Inf. Proc. Syst.* **34**, 29287 (2021).
- [14] P. Kantroo, G. P. Wagner, and B. B. Machta, Pseudo-perplexity in one fell swoop for protein fitness estimation, *PRX Life* (2025), doi:10.1103/zhx7-hcmm.
- [15] T. Bepler and B. Berger, Learning protein sequence embeddings using information from structure, in *International Conference on Learning Representations* (ICLR, 2019).
- [16] D. Sgarbossa, C. Malbranke, and A.-F. Bitbol, ProtMamba: A homology-aware but alignment-free protein state space model, [bioRxiv: 2024.05.24.595730](https://arxiv.org/abs/2024.05.24.595730).
- [17] S. Alamdari, N. Thakkar, R. van den Berg, A. X. Lu, N. Fusi, A. P. Amini, and K. K. Yang, Protein generation with evolutionary diffusion: Sequence is all you need, [bioRxiv: 2023.09.11.556673](https://arxiv.org/abs/2023.09.11.556673).
- [18] X. Wang, Z. Zheng, F. Ye, D. Xue, S. Huang, and Q. Gu, Diffusion language models are versatile protein learners, in

- Proceedings of the 41st International Conference on Machine Learning*, Vol. 235 (PMLR, 2024), pp. 52309–52333.
- [19] R. M. Rao, J. Liu, R. Verkuil, J. Meier, J. Canny, P. Abbeel, T. Sercu, and A. Rives, MSA transformer, in *Proceedings of the 38th International Conference on Machine Learning*, edited by M. Meila and T. Zhang, Proceedings of Machine Learning Research Vol. 139 (PMLR, 2021), pp. 8844–8856.
  - [20] J. Jumper, R. Evans, A. Pritzel, T. Green, M. Figurnov, O. Ronneberger, K. Tunyasuvunakool, R. Bates, A. Židek, A. Potapenko, A. Bridgland, C. Meyer, S. A. A. Kohl, A. J. Ballard, A. Cowie, B. Romera-Paredes, S. Nikolov, R. Jain, J. Adler, T. Back, S. Petersen, D. Reiman, E. Clancy, M. Zielinski, M. Steinegger, M. Pacholska, T. Berghammer, S. Bodenstein, D. Silver, O. Vinyals, A. W. Senior, K. Kavukcuoglu, P. Kohli, and D. Hassabis, Highly accurate protein structure prediction with Alpha Fold, *Nature (London)* **596**, 583 (2021).
  - [21] Z. Lin, H. Akin, R. Rao, B. Hie, Z. Zhu, W. Lu, N. Smetanin, R. Verkuil, O. Kabeli, Y. Shmueli, A. dos Santos Costa, M. Fazel-Zarandi, T. Sercu, S. Candido, and A. Rives, Evolutionary-scale prediction of atomic-level protein structure with a language model, *Science* **379**, 1123 (2023).
  - [22] A. Elofsson, Progress at protein structure prediction, as seen in CASP15, *Curr. Opin. Struct. Biol.* **80**, 102594 (2023).
  - [23] J. D. Thompson, B. Linard, O. Lecompte, and O. Poch, A comprehensive benchmark study of multiple sequence alignment methods: current challenges and future perspectives, *PLoS ONE* **6**, e18093 (2011).
  - [24] T. F. Truong and T. Bepler, PoET: A generative model of protein families as sequences-of-sequences, in *Proceedings of the 37th International Conference on Neural Information Processing Systems* (Curran Associates, Red Hook, NY, 2023).
  - [25] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W.-T. Yih, T. Rocktäschel, S. Riedel, and D. Kiela, Retrieval-augmented generation for knowledge-intensive nlp tasks, in *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20 (Curran Associates, Red Hook, NY, 2020).
  - [26] K. Guu, K. Lee, Z. Tung, P. Pasupat, and M.-W. Chang, Retrieval augmented language model pre-training, in *Proceedings of the 37th International Conference on Machine Learning*, ICML'20, Vol. 119 (PMLR, 2020), pp. 3929–3938.
  - [27] Z. Zheng, Y. Deng, D. Xue, Y. Zhou, F. Ye, and Q. Gu, Structure-informed language models are protein designers, in *Proceedings of the 40th International Conference on Machine Learning*, ICML'23, Vol. 202 (PMLR, 2023), pp. 42317–42338.
  - [28] See Supplemental Material at <http://link.aps.org/supplemental/10.1103/db1b-hy16> for additional information on training, ablations, sequence generation method, and motif scaffolding model comparisons.
  - [29] E. Nijkamp, J. A. Ruffolo, E. N. Weinstein, N. Naik, and A. Madani, Progen2: exploring the boundaries of protein language models, *Cell Syst.* **14**, 968 (2023).
  - [30] T. Hayes, R. Rao, H. Akin, N. J. Sofroniew, D. Oktay, Z. Lin, R. Verkuil, V. Q. Tran, J. Deaton, M. Wiggert, R. Badkundri, I. Shafkat, J. Gong, A. Derry, R. S. Molina, N. Thomas, Y. Khan, C. Mishra, C. Kim, L. J. Bartie, M. Nemeth, P. D. Hsu, T. Sercu, S. Candido, and A. Rives, Simulating 500 million years of evolution with a language model, *Science* **387**, 850 (2025).
  - [31] N. Bhattacharya, N. Thomas, R. Rao, J. Dauparas, P. K. Koo, D. Baker, Y. S. Song, and S. Ovchinnikov, Interpreting Potts and transformer protein models through the lens of simplified attention, *Biocomputing 2022* (World Scientific, Singapore, 2022), pp. 34–45.
  - [32] S. S. Sahoo, M. Arriola, Y. Schiff, A. Gokaslan, E. Marroquin, J. T. Chiu, A. Rush, and V. Kuleshov, Simple and effective masked diffusion language models, in *The Thirty-eighth Annual Conference on Neural Information Processing Systems* (NeurIPS, 2024).
  - [33] M. Mirdita, L. von den Driesch, C. Galiez, M. J. Martin, J. Söding, and M. Steinegger, Uniclust databases of clustered and deeply annotated protein sequences and alignments, *Nucl. Acids Res.* **45**, D170 (2017).
  - [34] G. Ahdriz, N. Bouatta, S. Kadyan, L. Jarosch, D. Berenberg, I. Fisk, A. M. Watkins, S. Ra, R. Bonneau, and M. AlQuraishi, OpenProteinSet: Training data for structural biology at scale, in *Proceedings of the 37th International Conference on Neural Information Processing Systems Datasets and Benchmarks Track* (Curran Associates, Red Hook, NY, 2023).
  - [35] S. Petti, N. Bhattacharya, R. Rao, J. Dauparas, N. Thomas, J. Zhou, A. M. Rush, P. K. Koo, and S. Ovchinnikov, End-to-end learning of multiple sequence alignments with differentiable Smith-Waterman, *Bioinformatics* **39**, btac724 (2023).
  - [36] U. Lupo, D. Sgarbossa, and A.-F. Bitbol, Protein language models trained on multiple sequence alignments learn phylogenetic relationships, *Nat. Commun.* **13**, 6298 (2022).
  - [37] Y. Zhang and J. Skolnick, Scoring function for automated assessment of protein structure template quality, *Proteins: Structure*, *Proteins* **57**, 702 (2004).
  - [38] S. R. Eddy, Accelerated profile HMM searches, *PLoS Comput Biol.* **7**, e1002195 (2011).
  - [39] J. Dauparas, I. Anishchenko, N. Bennett, H. Bai, R. J. Ragotte, L. F. Milles, B. I. M. Wicky, A. Courbet, R. J. de Haas, N. Bethel, P. J. Y. Leung, T. F. Huddy, S. Pellock, D. Tischer, F. Chan, B. Koepnick, H. Nguyen, A. Kang, B. Sankaran, A. K. Bera, N. P. King, and D. Baker, Robust deep learning-based protein sequence design using ProteinMPNN, *Science* **378**, 49 (2022).
  - [40] J. L. Watson, D. Juergens, N. R. Bennett, B. L. Trippe, J. Yim, H. E. Eisenach, W. Ahern, A. J. Borst, R. J. Ragotte, L. F. Milles, B. I. M. Wicky, N. Hanikel, S. J. Pellock, A. Courbet, W. Sheffler, J. Wang, P. Venkatesh, I. Sappington, S. V. Torres, A. Lauko, V. D. Bortoli, E. Mathieu, S. Ovchinnikov, R. Barzilay, T. S. Jaakkola, F. DiMaio, M. Baek, and D. Baker, De novo design of protein structure and function with RFdiffusion, *Nature (London)* **620**, 1089 (2023).
  - [41] R. Wu, F. Ding, R. Wang, R. Shen, X. Zhang, S. Luo, C. Su, Z. Wu, Q. Xie, B. Berger, J. Ma, and J. Peng, High-resolution de novo structure prediction from primary sequence, *bioRxiv*: 2022.07.21.500999.
  - [42] X. Wang, Z. Zheng, F. Ye, D. Xue, S. Huang, and Q. Gu, DPLM-2: A multimodal diffusion protein language model, in *The Thirteenth International Conference on Learning Representations* (ICLR, 2025).
  - [43] G. Munsamy, R. Illanes-Vicioso, S. Funcillo, I. T. Nakou, S. Lindner, G. Ayres, L. S. Sheehan, S. Moss, U. Eckhard, P. Lorenz, and N. Ferruz, Conditional language models enable the efficient design of proficient enzymes, *bioRxiv*: 2024.05.03.592223.
  - [44] J. Su, C. Han, Y. Zhou, J. Shan, X. Zhou, and F. Yuan, SaProt: Protein language modeling with structure-aware

- vocabulary, in *The Twelfth International Conference on Learning Representations* (ICLR, 2024).
- [45] M. Heinzinger, K. Weissenow, J. G. Sanchez, A. Henkel, M. Mirdita, M. Steinegger, and B. Rost, Bilingual language model for protein sequence and structure, *NAR Genom. Bioinf.* **6**, lqae150 (2024).
- [46] K. Michalewicz, M. Barahona, and B. Bravi, Integrating protein sequence embeddings with structure via graph-based deep learning for the prediction of single-residue properties, [arXiv:2502.17294](https://arxiv.org/abs/2502.17294).
- [47] D. Sgarbossa and A.-F. Bitbol, <https://github.com/Bitbol-Lab/rag-esm> (2025).