Generative power of a protein language model trained on multiple sequence alignments

Damiano Sgarbossa [1 2] **Umberto Lupo** [1 2] **Anne-Florence Bitbol** [1 2]

Abstract

Computational models starting from large ensembles of evolutionarily related protein sequences capture a representation of protein families and learn constraints associated to protein structure and function. They thus open the possibility for generating novel sequences belonging to protein families. Protein language models trained on multiple sequence alignments, such as MSA Transformer, are highly attractive candidates to this end. We propose and test an iterative method that directly uses the masked language modeling objective to generate sequences using MSA Transformer. We demonstrate that the resulting sequences generally score better than those generated by Potts models, and even than natural sequences, for homology, coevolution and structure-based measures. They also reproduce well the statistics and the distribution of sequences in sequence space of natural data.

1. Introduction

Designing new proteins with specific structure and function is a highly important goal of bioengineering. Indeed, it can allow to tune their stability or their biochemical properties, including their enzymatic activities, enabling important medical applications. *De novo* or rational protein design, which starts with target three-dimensional structures and physico-chemical potentials, can generate proteins which are not in a known family, but is restricted to small proteins. Conversely, directed evolution allows for a local search of sequence space, but remains limited to the vicinity of a natural sequence.

Generative computational models that build on the breadth of available natural protein sequence data, and capture a representation of protein families, now offer great alternatives that can allow to sample novel sequences belonging to protein families. In particular, Potts models (Figliuzzi et al., 2018), and variational autoencoders (Hawkins-Hooker et al., 2021a) have been shown to produce functional proteins.

Protein language models (PLMs) are deep learning models based on natural language processing methods, especially attention and transformers. They are trained on large ensembles of protein sequences, and capture long-range dependencies within a protein sequence (Elnaggar et al., 2021; Rives et al., 2021). They are able to predict structure from a single sequence in an unsupervised way (Rao et al., 2021a;b). The great success of supervised protein structure prediction by AlphaFold (Jumper et al., 2021) is partly based on the use of transformers. It is therefore of strong interest to assess the generative ability of PLMs, and recent works show that this has high potential (Madani et al.; Johnson et al., 2021; Ferruz et al., 2022).

Correlations in amino-acid usage between the columns of multiple sequence alignments (MSAs) of homologous proteins are important to generate functional synthetic proteins, and the success of Potts models relies on these correlations. PLMs that take MSAs as input are able to directly exploit this covariation signal, and are thus particularly interesting candidates for protein design. Thus motivated, we focus on MSA Transformer (Rao et al., 2021b), a PLM which was trained on MSAs using the masked language modeling objective, without additional supervised training – by contrast to AlphaFold. We ask how the generative properties of MSA Transformer compare to those of Boltzmann machine DCA (bmDCA) (Figliuzzi et al., 2018; Russ et al., 2020), a state-of-the-art generative Potts model.

2. Methods

We propose an *iterative masking procedure* (see Figure 1) that directly uses the masked language modeling objective iteratively to generate sequences using MSA Transformer.

Our algorithm, given an arbitrary MSA $\mathcal{M}$ of natural sequences, proceeds as follows:

[1] School of Life Sciences, Institute of Bioengineering, École Polytechnique Fédérale de Lausanne (EPFL), CH-1015, Lausanne, Switzerland [2] SIB Swiss Institute of Bioinformatics, CH-1015, Lausanne, Switzerland. Correspondence to: Umberto Lupo <umberto.lupo@epfl.ch>, Anne-Florence Bitbol <anne-florence.bitbol@epfl.ch>.

The 2022 ICML Workshop on Computational Biology. Baltimore, Maryland, USA, 2022. Copyright 2022 by the author(s).

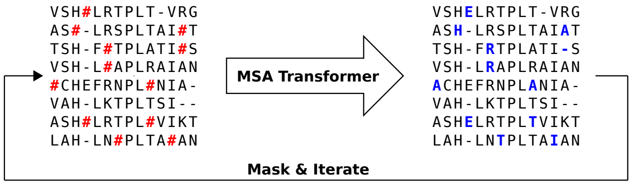

Figure 1. **Iterative masking procedure to generate sequences using MSA Transformer.** We randomly mask each residue in the input MSA with masking probability $p = 0.1$, use the model to fill in masked entries, and iterate.

1. If necessary (for memory reasons), sample some sequences (i.e. rows) from $\mathcal{M}$ to obtain an input MSA $\mathcal{M}'$ of sequences picked uniformly at random.

2. Randomly mask each residue of $\mathcal{M}'$ with masking probability $p = 0.1$, otherwise leave it unchanged.

3. Feed the masked MSA to MSA Transformer, and fill each masked entry with the token with highest probability (obtained from the output logits).

4. Repeat points **2-3** for $I = 200$ iterations.

For each natural MSA $\mathcal{M}$, we repeat the procedure above multiple times, sampling sequences each time from $\mathcal{M}$ without replacement to obtain a different input MSA $\mathcal{M}'$, until all the sequences in $\mathcal{M}$ are used. Note that sequences remain aligned at all times during the procedure. Combining the MSAs resulting from all these batches then yields a synthetic MSA with the same depth as the natural one.

The values of p and I were chosen so that the properties of the synthetic MSA have converged. Details are described in our preprint (Sgarbossa et al.).

Datasets. To generate synthetic MSAs with MSA Transformer and bmDCA and compare them to their natural counterparts, we consider the deep Pfam "full" alignments associated to 14 protein domains (Figliuzzi et al., 2018).

3. Results

An iterative masking procedure allows MSA Transformer to generate novel sequences with high scores. Is MSA Transformer able to generate sequences that are credible members of protein families? How do its generative abilities compare to bmDCA, a state-of-the-art generative Potts model which has been experimentally shown to generate functional proteins (Russ et al., 2020)? To address these questions, we employed an iterative masking procedure to generate synthetic MSAs from natural MSAs of 14 different Pfam protein families (see (Figliuzzi et al., 2018)) with MSA Transformer, as described in Section 2. We also generated synthetic sequences by MCMC sampling from Potts models inferred from these MSAs by bmDCA. For each protein family, we obtained three different MSAs of the same depth: the natural one, the one generated by our iterative masking procedure using MSA Transformer, and the one sampled from the inferred Potts model. To characterize each sequence, we consider three different scores. First, we assess the quality of the generated sequences as homologs of the protein family of interest via the HMMER (http://hmmer.org) score of the hidden Markov model employed by Pfam to retrieve natural homologs. Second, we consider a score that accounts for coevolution between amino-acid sites, namely the statistical energy score from the Potts model fitted on the natural MSA. Third, we determine AlphaFold's confidence in its determination of the three-dimensional structure of these sequences, via the predicted local-distance difference test (pLDDT) score. All scores are such that higher values are better. Each one accounts for a different aspect of proteins (scores described in detail in (Sgarbossa et al.)).

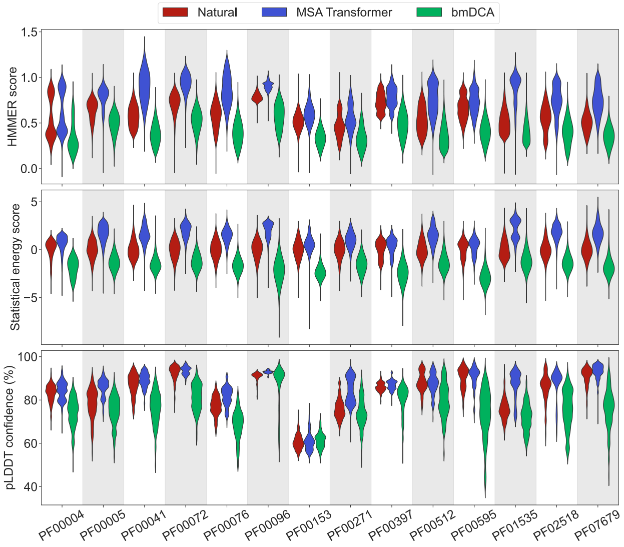

Figure 2. **Comparison of homology, coevolution, and structure-based scores between natural and generated sequences.** For each Pfam family in (Figliuzzi et al., 2018), we compare a natural MSA from Pfam and two synthetic MSAs of the same depth, respectively generated using MSA Transformer or bmDCA. For each of three scores, we show its distribution among sequences in each MSA. Higher score values are better. **Top:** For each Pfam family, HMMER scores are divided by the highest score found in the natural MSA. **Middle:** Statistical energy scores are defined as minus the bmDCA statistical energies (shifted by the mean of natural MSA scores, and normalized by their standard deviation). **Bottom:** AlphaFold's pLDDT confidence scores coming from 200 randomly chosen sequences from each MSA.

Figure 2 shows that, for all protein families considered, and for these three different scores, the sequences generated by MSA Transformer using our iterative masking procedure have scores that are comparable and even better, on average, than natural sequences. The opposite holds for sequences generated by bmDCA. We tested this in a quantitative way by employing the Kolmogorov-Smirnov test which confirms our observations. Thus, MSA Transformer is a good candidate to generate synthetic sequences from protein families via our iterative masking procedure.

Good scores are not restricted to synthetic sequences similar to natural ones. How different are these synthetic sequences from the natural ones? In particular, are those that score best original sequences, or almost copies of natural sequences? In Figure 3 we show, for one example protein family (PF00153), the HMMER score and the DCA statistical energy score versus the sequence's Hamming distance to its closest natural sequence in the natural MSA. Trends were similar for the other families we studied.

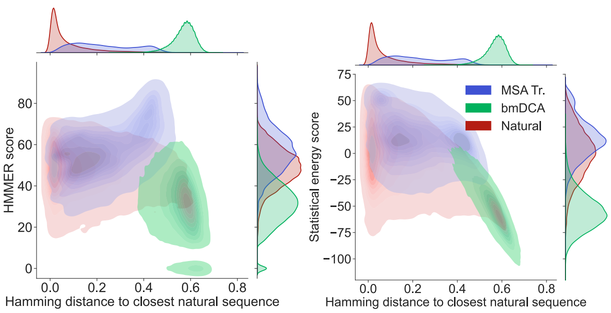

Figure 3. **Homology and coevolution scores vs. distance to the natural MSA, for protein family PF00153.** We show contour plots of the HMMER score and the statistical energy score (defined as minus the DCA statistical energy, shifted by its mean value in the natural MSA) versus the Hamming distance of each sequence to the closest natural sequence (which is not itself, in the case of natural sequences).

Figure 3 shows that MSA Transformer generates sequences with variable distances to their closest natural sequences, and that these distances are overall larger than those between closest natural sequences. By contrast, bmDCA generates sequences which are often extremely different from the natural ones. In addition, at a given distance from the natural MSA, the MSA Transformer-generated sequences often have higher HMMER scores and statistical energy scores than the bmDCA-generated ones. Furthermore, the MSA Transformer-generated sequences featuring the highest HMMER scores are those with large Hamming distances to natural sequences, i.e. truly original ones. Therefore, MSA Transformer-generated sequences are not reaching good scores by overfitting and reproducing natural sequences. The fraction of MSA Transformer-generated sequences which are identical to sequences in the input natural MSAs is below 0.05% for all families considered, except PF00595, PF00096 and PF00397 for which this fraction is $4, 5$ and 42%, respectively. These families feature low diversity and short length.

Quantitatively, in Figure 3, the Pearson correlation between HMMER scores and Hamming distances to closest natural sequences are respectively $\rho = 0.52$ and $\rho = -0.33$ for MSA Transformer and for bmDCA. Moreover, a strong negative correlation between the statistical energy score and the Hamming distance is observed for bmDCA, while it is much weaker for MSA Transformer. We observe these trends for most protein families studied, and also when using BLOSUM similarity scores instead of Hamming distances. Sequences generated by bmDCA were already reported to have overall worse statistical energy scores than their natural counterparts, and decreasing the sampling temperature below 1 was proposed as a mitigating strategy (Russ et al., 2020). However, we observed that this substantially decreased the fitting of first- and second-order statistics, and only slightly improved HMMER scores (Sgarbossa et al.).

MSA Transformer captures well the distribution of sequences in sequence space. How are synthetic MSAs generated by MSA Transformer and bmDCA impacted by the heterogeneous repartition of natural sequences in sequence space? While natural protein sequences in a family have evolved from a common ancestor along a phylogeny, synthetic sequences do not have a real evolutionary history. However, as bmDCA and MSA Transformer are trained on natural data, they can capture phylogenetic correlations (Lupo et al.). Besides, inferred Potts models are known to be impacted by phylogenetic correlations.

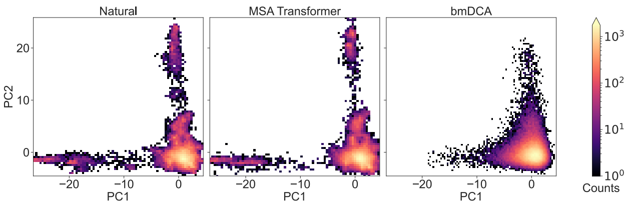

Figure 4. **Distribution of sequences in sequence space, for family PF00153.** We show the distribution of one-hot encoded natural and synthetic sequences projected in the subspace of the first two principal components of the natural MSA.

To analyze the distribution of MSA sequences in sequence space, we perform a principal component analysis of one-hot encoded MSAs, and focus on the top two principal components (PCs) of natural MSAs (Figliuzzi et al., 2018). Figure 4 shows the distribution of sequences in the space spanned by these top two PCs, for natural and synthetic

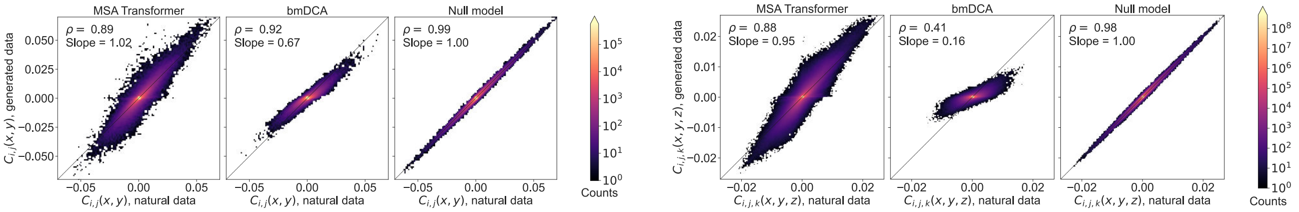

Figure 5. **Two- and three-body connected correlations estimated from generated MSAs versus the natural one, for family PF00153.** Relationships between connected correlations estimated from the MSA generated by MSA Transformer or bmDCA, and those estimated from the natural MSA, are shown as binned scatter plots both for two-body (top row) and three-body (bottom row) statistics. To assess finite-size effects, we include a null model (third column) obtained by splitting the natural MSA in half and comparing the statistics of one half with those of the other. Pearson correlation coefficients ρ, and slopes of lines of best fit, are reported in each case.

MSAs, in the case of PF00153. We observe that MSA Transformer is able to generate sequences with a distribution in sequence space that is very similar to the natural one, while bmDCA smoothes this distribution. This observation is general across all the MSAs we considered. Note however that the top two PCs explain a small fraction of the variance.

Higher-order statistics are better reproduced by MSA Transformer, while lower-order statistics are better reproduced by bmDCA. How well do synthetic MSAs generated by MSA Transformer and bmDCA reproduce the statistics of amino-acid usage observed in natural MSAs? To address this question, Figure 5 shows a comparison of second-order connected correlations for natural and synthetic MSAs of family PF00153: MSA Transformer reproduces these correlations less accurately than bmDCA. This makes sense because Potts models are pairwise maximum entropy models constrained to match the one- and two-body frequencies from natural MSAs, so bmDCA is trained to reproduce these frequencies. Thus, this result is in line with expectations. What about higher-order statistics which are not fitting objectives of bmDCA? In Figure 5, we show a comparison of third-order connected correlations for PF00153: the MSA generated by MSA Transformer reproduces the higher-order statistics of the natural MSA better than the one generated by bmDCA. These conclusions are quite generic across the 14 MSAs we considered.

4. Discussion

We proposed an iterative masking procedure which directly exploits the masked language modeling objective of PLMs to generate sequences using the MSA-based neural language model MSA Transformer. We found that these sequences generally score better than natural ones and that those generated by bmDCA Potts models on three very different aspects, namely homology, coevolution and structure-based scores. Moreover, MSA Transformer-generated sequences better reproduce the higher-order statistics and the distribution of sequences in sequence space of natural data than bmDCA-generated ones, which conversely better reproduces lower-order statistics, consistently with its training objective.

Our results are highly promising for sequence generation by MSA-based PLMs, and we hope that they will motivate further studies, especially experimental tests. More generally, our results reinforce the new promising “coevolution-driven” protein design method based on exploiting sequences of evolutionarily related proteins. This concept differs from structure- and physics-based *de novo* design, and from the new possibility to use supervised deep learning models able to accurately predict protein structures (Jumper et al., 2021; Baek et al., 2021; Chowdhury et al., 2021) for structure-driven sequence generation. The coevolution-driven approach is intermediate between structure-based approaches and directed evolution ones. It was recently experimentally validated in the cases of bmDCA Potts models (Russ et al., 2020) and variational autoencoders (Hawkins-Hooker et al., 2021a). PLMs trained on multiple sequence alignments provide state-of-the-art unsupervised contact prediction in tied row attentions, and capture phylogenetic relationships in column attentions (Lupo et al.). This makes them ideal candidates to generate new protein sequences from given families. However, contrary to Potts models and variational autoencoders, they do not allow direct sampling from a probability distribution over sequences. Here, we demonstrated the power of a simple generation method directly based on the masked language modeling objective. It differs from using a decoder in this context, which allows autoregressive generation of sequences, but requires training a full encoder-decoder model and learning a parametric function mapping an MSA to a distribution over its sequences (Hawkins-Hooker et al., 2021b). We instead directly employed the representation of protein families captured by the self-supervised model MSA Transformer to generate sequences. More sophisticated sampling methods could be considered (Goyal et al., 2021), but our minimal approach already gives very promising results. Starting from large MSA-free models (Bileschi et al., 2022; Shin et al., 2021; Madani et al.) is another promising direction.

References

Baek, M., DiMaio, F., Anishchenko, I., Dauparas, J., Ovchinnikov, S., Lee, G. R., Wang, J., Cong, Q., Kinch, L. N., Schaeffer, R. D., Millán, C., Park, H., Adams, C., Glassman, C. R., DeGiovanni, A., Pereira, J. H., Rodrigues, A. V., van Dijk, A. A., Ebrecht, A. C., Opperman, D. J., Sagmeister, T., Buhlheller, C., Pavkov-Keller, T., Rathinaswamy, M. K., Dalwadi, U., Yip, C. K., Burke, J. E., Garcia, K. C., Grishin, N. V., Adams, P. D., Read, R. J., and Baker, D. Accurate prediction of protein structures and interactions using a three-track neural network. *Science*, 373(6557):871–876, 2021.

Bileschi, M. L., Belanger, D., Bryant, D., Sanderson, T., Carter, B., Sculley, D., DePristo, M. A., and Colwell, L. J. Using deep learning to annotate the protein universe. *Nature Biotechnol.*, 2022. doi: 10.1101/626507.

Chowdhury, R., Bouatta, N., Biswas, S., Rochereau, C., Church, G. M., Sorger, P. K., and AlQuraishi, M. Single-sequence protein structure prediction using language models from deep learning. *bioRxiv*, 10.1101/2021.08.02.454840, 2021.

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 deep learning and high performance computing. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021.

Ferruz, N., Schmidt, S., and Höcker, B. A deep unsupervised language model for protein design. *bioRxiv*, 10.1101/2022.03.09.483666, 2022.

Figliuzzi, M., Barrat-Charlaix, P., and Weigt, M. How pairwise coevolutionary models capture the collective residue variability in proteins? *Mol. Biol. Evol.*, 35(4): 1018–1027, 2018.

Goyal, K., Dyer, C., and Berg-Kirkpatrick, T. Exposing the implicit energy networks behind masked language models via Metropolis–Hastings. *arXiv*, 10.48550/arxiv.2106.02736, 2021.

Hawkins-Hooker, A., Depardieu, F., Baur, S., Couairon, G., Chen, A., and Bikard, D. Generating functional protein variants with variational autoencoders. *PLoS Comput. Biol.*, 17(2):1–23, 02 2021a.

Hawkins-Hooker, A., Jones, D. T., and Paige, B. MSA-conditioned generative protein language models for fitness landscape modelling and design. In *Machine Learning for Structural Biology Workshop, NeurIPS*, 2021b.

Johnson, S. R., Massie, K., Monaco, S., and Syed, Z. Generating novel protein sequences using Gibbs sampling of masked language models. *bioRxiv*, 2021. doi: 10.1101/2021.01.26.428322.

Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., Bridgland, A., Meyer, C., Kohl, S. A. A., Ballard, A. J., Cowie, A., Romera-Paredes, B., Nikolov, S., Jain, R., Adler, J., Back, T., Petersen, S., Reiman, D., Clancy, E., Zielinski, M., Steinegger, M., Pacholska, M., Berghammer, T., Bodenstein, S., Silver, D., Vinyals, O., Senior, A. W., Kavukcuoglu, K., Kohli, P., and Hassabis, D. Highly accurate protein structure prediction with AlphaFold. *Nature*, 596:583–589, 2021.

Lupo, U., Sgarbossa, D., and Bitbol, A.-F. Protein language models trained on multiple sequence alignments learn phylogenetic relationships. *bioRxiv*, 10.1101/2022.03.29.486219.

Madani, A., Krause, B., Greene, E. R., Subramanian, S., Mohr, B. P., Holton, J. M., Olmos, J. L., Xiong, C., Sun, Z. Z., Socher, R., Fraser, J. S., and Naik, N. Deep neural language modeling enables functional protein generation across families. *bioRxiv*, 10.1101/2021.07.18.452833.

Rao, R., Meier, J., Sercu, T., Ovchinnikov, S., and Rives, A. Transformer protein language models are unsupervised structure learners. In *9th International Conference on Learning Representations, ICLR 2021*, 2021a.

Rao, R. M., Liu, J., Verkuil, R., Meier, J., Canny, J., Abbeel, P., Sercu, T., and Rives, A. MSA Transformer. *Proceedings of the 38th International Conference on Machine Learning*, 139:8844–8856, 18–24 Jul 2021b.

Rives, A., Meier, J., Sercu, T., Goyal, S., Lin, Z., Liu, 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. *Proc. Natl. Acad. Sci. U.S.A.*, 118(15), 2021.

Russ, W. P., Figliuzzi, M., Stocker, C., Barrat-Charlaix, P., Socolich, M., Kast, P., Hilvert, D., Monasson, R., Cocco, S., Weigt, M., and Ranganathan, R. An evolution-based model for designing chorismate mutase enzymes. *Science*, 369(6502):440–445, 2020.

Sgarbossa, D., Lupo, U., and Bitbol, A.-F. Generative power of a protein language model trained on multiple sequence alignments. *bioRxiv*. doi: 10.1101/2022.04.14.488405.

Shin, J. E., Riesselman, A. J., Kollasch, A. W., McMahon, C., Simon, E., Sander, C., Manglik, A., Kruse, A. C., and Marks, D. S. Protein design and variant prediction using autoregressive generative models. *Nature Commun.*, 12 (1):2403, 04 2021.