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,*]

1 Institute of Bioengineering, School of Life Sciences, École Polytechnique Fédérale de Lausanne (EPFL), CH-1015 Lausanne, Switzerland
2 SIB Swiss Institute of Bioinformatics, CH-1015 Lausanne, Switzerland
* Email: umberto.lupo@epfl.ch, anne-florence.bitbol@epfl.ch

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. Moreover, MSA Transformer better reproduces the higher-order statistics and the distribution of sequences in sequence space of natural data than Potts models, although Potts models better reproduce first- and second-order statistics. MSA Transformer is thus a strong candidate for protein sequence generation and protein design.

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. The search for novel proteins is difficult due to the huge size of protein sequence space: for instance, there are 20^{100} different possible sequences for a short protein domain with 100 amino acids. Furthermore, only a small fraction of this space comprises sequences that do fold, as demonstrated by experiments studying random sequences [1], and by theoretical arguments based on the physics of disordered systems [2]. *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 protein family [3–5], but is generally restricted to small proteins [6]. Conversely, directed evolution allows to perform a local search of sequence space, but generally remains limited to the vicinity of a natural sequence [7].

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, or DCA models [8–11], which are pairwise maximum entropy models trained to reproduced

the one- and two-body statistics of the sequences of a family, allow direct sampling from a probability distribution modeling this family [12], and have been used successfully for protein design [13]. Variational autoencoders are deep learning models which also allow sampling, and they have been shown to successfully produce functional proteins [14], although their statistical properties appear to have a lower quality than with Potts models [15].

Protein language models are deep learning models based on natural language processing methods, especially attention [16] and transformers [17]. They are trained on large ensembles of protein sequences, and capture long-range dependencies within a protein sequence [18–25]. They are able to predict structure from a single sequence in an unsupervised way [21, 25], potentially by transferring knowledge from their large training set [26, 27]. The great success of supervised protein structure prediction by AlphaFold [28] is partly based on the use of transformers. It is therefore of strong interest to assess the generative ability of protein language models, and recent works show that this has high potential [24, 29–32].

Correlations in amino-acid usage that can be observed between the columns of multiple sequence alignments (MSAs) of homologous proteins [33–35] were experimentally demonstrated to be highly important to generate functional synthetic proteins [1, 36]. The importance of pairwise coevolution signals was then corroborated by the success of Potts models at predicting structural contacts [8–11, 37], analyzing mutational effects [38–41], protein evolution [42] and conformational changes [43, 44], designing proteins [13], and at predicting protein-protein interaction partners [45–48]. Protein language models that take MSAs as input [25, 28] are able to directly exploit this covariation signal, and are thus particularly interesting candidates for protein design. Thus motivated, we focus on MSA Transformer [25], a protein language model which was trained on MSAs using the masked language objective, without additional supervised training – by contrast to AlphaFold [28]. We ask how the generative properties of MSA Transformer compare to those of Boltzmann machine DCA (bmDCA) [12, 13], a state-of-the-art generative Potts model.

We propose and test a generating method that directly uses the masked language objective in an iterative way to generate sequences using MSA Transformer. Using homology, coevolution and structural scores, we demonstrate that the sequences generated by this method score better than those generated by bmDCA and even than natural sequences. We further show that this good performance is not restricted to synthetic sequences that are very similar to natural sequences. However, we find that bmDCA better reproduces the one- and two-body statistics of the natural MSAs than MSA Transformer, consistently with its training objective. Interestingly, the opposite generally holds for three-body statistics. MSA-Transformer–generated sequences also better reproduce the distribution of sequences in sequence space than bmDCA-generated ones. Our conclusion is that MSA Transformer is a strong candidate for protein sequence generation and protein design.

Methods

Using MSA Transformer to generate sequences via an iterative masking procedure

Iterative masking procedure. In order to generate new sequences using MSA Transformer, we directly leverage the model's ability to assign, to arbitrary masked residue positions, a probability for each of the possible amino-acid tokens, given by the softmax of the model's output logits [21, 49, 50]. Indeed, in its pre-training, MSA Transformer applies the masked language modeling (MLM) objective to a training set of 26 million MSAs [21]. For this, it minimizes a pseudolikelihood loss, which reads, for an MSA $\mathcal{M}$, and a version $\widetilde{\mathcal{M}}$ of $\mathcal{M}$ in

which some amino acids (those in a "mask") are masked:

$$\mathcal{L}_{\text{MLM}}(\mathcal{M}, \widetilde{\mathcal{M}}; \theta) = -\sum_{(m,i) \in \text{mask}} \log p(x_{m,i} \,|\, \widetilde{\mathcal{M}}; \theta)\,. \tag{1}$$

Here, $x_{m,i}$ denotes the amino acid at the i-th residue position in the m-th sequence of $\mathcal{M}$, and θ denotes all model parameters. For each position i in each sequence m, the model outputs one value ("logit") per amino-acid/gap symbol, and softmax-normalizing all values from this location in the MSA yields the conditional probabilities $p(x_{m,i} \,|\, \widetilde{\mathcal{M}}; \theta)$ in Eq. (1), which are then summed over the subset of masked MSA locations.

We propose an iterative masking procedure (see Fig. 1) which, given an arbitrary MSA $\mathcal{M}$ of natural sequences, proceeds as follows:

1. If necessary, subsample $\mathcal{M}$ to obtain an input MSA $\mathcal{M}'$ for MSA Transformer. The depth of $\mathcal{M}'$ is chosen given the memory footprint of MSA Transformer. In practice, we use input MSAs containing 600 sequences,[1] picked uniformly at random from our natural MSAs.

2. Randomly mask each residue of $\mathcal{M}'$ with a masking probability p. In practice, we choose $p = 0.1$.

3. Feed the masked MSA to the model, and fill each masked entry with the token with highest probability.

4. Repeat Steps 2 and 3 a number of times. In practice, we stop the algorithm after $I = 200$ iterations.

As natural MSAs, we use Pfam full MSAs for 14 protein families, described in "Datasets". For each natural MSA $\mathcal{M}$, we repeat the procedure above multiple times, sampling each time from $\mathcal{M}$ without replacement to obtain a different input MSA $\mathcal{M}'$ in Step 1, until all the sequences in $\mathcal{M}$ are used. Combining the MSAs resulting from all these batches then yields a synthetic MSA with the same depth as the natural one, which ensures that the statistical properties of the synthetic MSA are subject to the same magnitude of finite-size errors as those of the natural MSA.

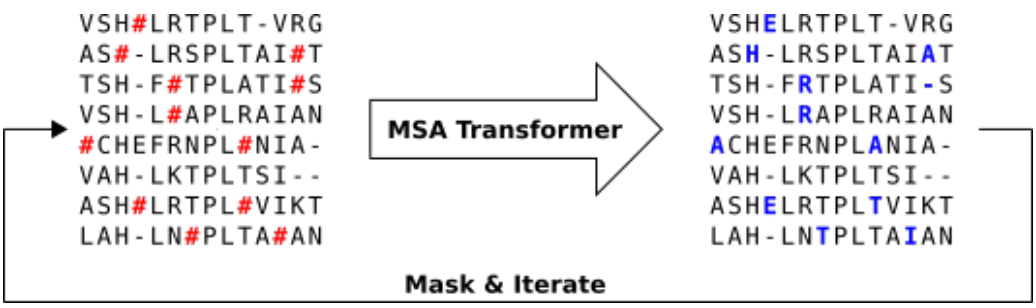

Figure 1: Iterative masking procedure to generate sequences using MSA Transformer.

Choosing parameters in the iterative masking procedure. Fig. S1 illustrates, in the case of Pfam family PF00153 and for different values of the masking probability p, how different properties of the generated MSAs evolve with the number I of iterations in the iterative masking procedure. These behaviors are generic across the protein families we studied. For

[1] During training, the authors of Ref. [25] kept $LM < 2^{14}$, where L is sequence length and M is MSA depth. However, we found that during inference we can use 2^{17} tokens on an Nvidia V100 32 GB GPU.

$p < 0.5$, we observe a gradual divergence from the initial natural sequences (Fig. S1A-B) and a simultaneous increase of scores (Fig. S1C-D, see "Scoring individual sequences" for definitions) and decrease of MSA diversity (Fig. S1E), and then a saturation of these various measures, as I increases. Our choice $I = 200$ is motivated by the fact that plateaus are reached at this point. However, the final values of all scores depend on p. Fig. S2 shows the contact maps inferred by MSA Transformer (using the logistic regression on tied row attentions trained in Ref. [25]) from generated sequences, for various values of I and p. We observe that the contact map characteristic of the protein family of interest gets gradually lost as I is increased for larger values of p (see Fig. S1F and Fig. S2)). These issues when p is increased are understandable, given that the pseudolikelihood loss used for the MLM objective in MSA Transformer ignores dependencies between masked entries. We note that despite this, larger values of p yield overall better average HMMER scores [51] and statistical energy scores (for $p < 0.5$). Our choice of $p = 0.1$ is motivated by the fact that this value is close to that employed in the training of the model ($p \approx 0.12$) [25], and that it better preserves contact maps. The product pI gives the average number of times that each amino acid of the MSA is changed during the generation process. With our choices, each amino acid is masked 20 times on average.

Variants of the iterative masking procedure. In our algorithm, we mask tokens randomly throughout the input MSA. We also explored an alternative procedure where masking is restricted to the first sequence of the input MSA. Thus, all other sequences act as a context for the first sequence which is gradually modified. This can be done either with a fixed context, or by sampling different sequences from the natural MSA at each iteration to form a variable context. Note that the procedure with fixed context is reminiscent of the non-iterative one used in Ref. [52] to compute deep mutational scanning scores from MSA Transformer. For the same masking probability $p = 0.1$ as in our standard procedure (note that fewer iterations are needed for convergence, in practice $I = 20$ suffices), the alternative procedure with fixed context yields sequences that are overall slightly less different from natural ones than the standard iterative masking procedure, while the opposite holds with variable context. Besides, both alternative procedures yield sequences with better HMMER scores, but worse statistical energy scores, than natural ones – see Table S1. Finally, the two- and three-body statistics (defined in "Analyzing the statistics of MSAs") of the natural MSA are less well reproduced using these alternative procedures than the standard one – see Table S1. We also note that these variants are computationally more demanding. In this context, we decided to focus on the standard iterative masking procedure.

There are also different ways of selecting the token to fill each masked position. We have chosen a greedy sampling method where the token with highest probability is selected. We also explored an alternative method where the new token to fill the masked position is chosen by sampling the probability distribution given by the softmax of the logits, see Eq. (1). This method allows to introduce a sampling temperature T into the softmax operation and compute the probability as $p = \mathrm{softmax}(\boldsymbol{\xi}/T)$, where $\boldsymbol{\xi}$ is the logit vector. Note that the greedy method that we employ corresponds to sampling at $T = 0$. We found that MSAs generated with higher values of T are farther from the corresponding natural MSAs, showing that increasing this sampling temperature promotes originality. However, they are of lower quality according to our HMMER and statistical energy scores, and reproduce the statistics of the natural data less well. These results, summarized in Table S1, motivated us to mainly consider greedy sampling.

Finally, in our iterative masking procedure, we subsample the initial natural MSAs uniformly at random. We also tried diversity maximizing sampling [25], but we found that random sampling gives slightly better results.

Sampling sequences from Potts models

To sample independent equilibrium sequences from Potts models, we used the strategy described in Ref. [53]. Specifically, we fitted Potts models on each of our natural MSAs using bmDCA [12] (https://github.com/ranganathanlab/bmDCA) with its default hyperparameters. Using bmDCA is known to yield Potts models with good generative power [12, 13].

Consider a sequence of L amino-acid sites. We denote by $x_i \in \{1, \ldots, q\}$ the state of site $i \in \{1, \ldots, L\}$, where $q = 21$ is the number of possible states, namely the 20 natural amino acids and the alignment gap. The Potts model Hamiltonian of a sequence $\boldsymbol{x} = (x_1, \ldots, x_L)$ reads [8, 54]:

$$H(\boldsymbol{x}) = -\sum_{i=1}^{L} h_i(x_i) - \sum_{j=1}^{L} \sum_{i=1}^{j-1} e_{ij}(x_i, x_j) \,. \tag{2}$$

For each MSA $\mathcal{M}$ in Table 1, we inferred parameters $h_i(x_i)$ and $e_{ij}(x_i, x_j)$ by bmDCA [12, 13]. The Potts model probability distribution is then given by the Boltzmann distribution associated to the Hamiltonian H in Eq. (2):

$$P(\boldsymbol{x}) = \frac{e^{-H(\boldsymbol{x})/T}}{Z} \,, \tag{3}$$

where Z is a constant ensuring normalization and T is a parameter whose default value is 1. To generate a synthetic MSA from $\mathcal{M}$, we performed equilibrium Markov Chain Monte Carlo (MCMC) sampling from the Potts model with Hamiltonian H in Eq. (2). Specifically, we used the implementation in Ref. [53] of the Metropolis–Hastings algorithm, in which each step is a proposed mutation at a single amino-acid site. We started from a set of M randomly and independently initialized sequences, where M is the depth of $\mathcal{M}$, and made a total number N of Monte Carlo steps on each sequence. For each $\mathcal{M}$, suitable values for N are estimated by bmDCA during its training, to ensure that Metropolis–Hastings sampling reaches thermal equilibrium after N steps when starting from a randomly initialized sequence [12]. We thus used the value of N estimated by bmDCA at the end of training. This yielded, for each MSA in Table 1, a synthetic MSA of the same depth, composed of independent equilibrium sequences.

This procedure allows to tune the sampling temperature T, in a similar spirit as for MSA Transformer, c.f. "Variants of the iterative masking procedure". This amounts to tuning the selection strength. Recall that Potts models are inferred at $T = 1$, which is thus the default value. Using MCMC sampling as described above, we generated synthetic MSAs at $T = 1$. We also considered lower values ($T = 0.9$ and $T = 0.95$), as in Ref. [13]. While decreasing T improves statistical energy scores, it substantially impairs the fitting of the one- and two-body statistics, decreases the Hamming distances to the closest natural sequences, and only yields a minor increase of the HMMER score, as shown in Table S1. Therefore, we focused on sequences generated at $T = 1$.

Scoring individual sequences

We use different scores to compare natural and generated sequences.

First, HMMER scores [51] are computed, for each sequence, from the Pfam profile Hidden Markov Models (HMM), employing the function `hmmsearch` from the HMMER Suite version 3.3.2 (http://hmmer.org). HMMER scores are homology scores, which are in particular used in Pfam to search sequence databases for sequence homologs and to construct full MSAs starting from curated seed MSAs. Higher HMMER scores indicate better matches to the Pfam HMM.

Second, DCA statistical energy scores are computed for each sequence using the Potts model Hamiltonian H in Eq. (2) with the couplings and the fields inferred by bmDCA on the natural MSA of the family of interest (see "Sampling sequences from Potts models"). The statistical

energy score is then defined as the opposite of the statistical energy, i.e. $-H(\boldsymbol{x})$ for a sequence $\boldsymbol{x}$, so that, here too, higher values mean better scores.

We also compute AlphaFold [28] structural prediction confidence scores, i.e. predicted local-distance difference test (pLDDT) values. Given the computational cost, for each natural or generated MSA, we evaluate pLDDT values for a subset of 200 randomly sampled sequences. Because AlphaFold takes MSAs as input, we compute these scores using the whole natural MSA of the family of interest as context in all cases.

Analyzing the statistics of MSAs

To compare the generated MSAs to the natural ones, we consider different statistical measures and information theory generalizations of these measures.

In each MSA, we first compute the one-body frequencies of occurrence of each amino acid at each site, the two-body frequencies of each pair of amino acids at each pair of sites, and the three-body frequencies associated to triplets. We denote them by $f_i(x)$, $f_{ij}(x,y)$, $f_{ijk}(x,y,z)$, where i, j and k denote sites, while x, y and z represent amino acids (see "Sampling sequences from Potts models"). We then estimate the second and third order connected correlations as:

$$C_{ij}(x,y) = f_{ij}(x,y) - f_i(x)f_j(y)\,; \tag{4}$$

$$\begin{aligned} C_{ijk}(x,y,z) = {} & f_{ijk}(x,y,z) - f_{ij}(x,y)f_k(z) - f_{ik}(x,z)f_j(y) - f_{jk}(y,z)f_i(x) \\ & + 2f_i(x)f_j(y)f_k(z)\,. \end{aligned} \tag{5}$$

We also compute the "plug-in" estimates of the Shannon entropy of each site H_i, and of the two- and three-body joint entropies H_{ij} and H_{ijk}, from the frequencies. They yield the plug-in estimates of the mutual information I_{ij} between two columns, and of the *co-information* I_{ijk} between three columns:

$$I_{ij} = H_i + H_j - H_{ij}\,; \tag{6}$$

$$I_{ijk} = H_i + H_j + H_k - H_{ij} - H_{ik} - H_{jk} + H_{ijk}\,. \tag{7}$$

Co-information is a measure of higher-order statistical dependencies [55–58], which generalizes mutual information to triplets of random variables, vanishes for independent variables, and reflects the balance between redundancy and synergy in these triplets [59, 60]. A systematic finite-size error occurs when estimating entropies using the plug-in estimate from frequencies measured in finite datasets [2], and it affects entropy-derived quantities such as mutual information and co-information. Here, we will not attempt to correct it. Rather, we will only make comparisons between MSAs of the same length and depth, which are affected by the same finite-size errors.

Characterizing the distribution of sequences in MSAs

Another way of studying the properties of generated MSAs is to analyze the distribution of their sequences in sequence space, and to compare it to that of natural sequences in the same family.

First, to assess whether generated sequences most resemble natural ones that are well-represented in their family or, rather, rare ones, we consider for each synthetic sequence its closest natural sequence. We then count the number of neighbors of this natural sequence in the natural MSA, i.e. the number of natural sequences that have (normalized) Hamming distance below $\delta = 0.2$ with the sequence of interest.[2]

[2]Note that the inverse of this number of neighbors gives the sequence weight w_i introduced in Eq. (8).

Second, to explore the distributions in sequence space of sequences within each MSA, and compare synthetic and natural MSAs, we associate to each sequence the concatenation of the one-hot encodings of each of its amino acids [12]. We perform a principal component analysis of the matrix corresponding to the natural MSA in this representation. We can then represent natural and synthetic sequences as points projected in the space defined by the first two principal components of the natural MSA.

Third, to analyze in more detail the apparent relatedness of generated sequences, and compare it to real phylogenetic relationships in natural sequences, we infer phylogenetic trees from each MSA using `FastTree 2` [61]. To quantitatively compare the topologies of these trees, which do not have the same leaves, we analyze the eigenvalue spectrum of their modified graph Laplacian (MGL) [62]. The MGL of a phylogenetic tree is defined as the difference between its degree matrix (a diagonal matrix whose i-th diagonal entry is the sum of the branch lengths from node i to all other nodes in the tree) and the matrix of patristic distances (whose (i, j)-th entry is the branch length between nodes i and j). Given the computational cost of running such an analysis on our deep MSAs, we use a bootstrap-aggregating strategy in the spirit of Ref. [63]. Namely, for each MSA we compute 200 different trees, each one inferred from a different sub-MSA of 500 sequences, itself randomly sampled from the whole MSA. Then, for each of these trees, we compute the eigenvalue spectrum of the MGL. Next, we merge all these spectra together to obtain a single eigenvalue spectral density. Note that this method has the advantage of not depending on the details of the topology of one large inferred tree, which are known to be sensitive to the choice of phylogeny reconstruction algorithm.

Datasets

To generate synthetic MSAs with MSA Transformer and bmDCA and compare them to their natural counterparts, we consider the deep Pfam "full" alignments [64] associated to 14 different protein domains (Table 1). Each MSA is a matrix $\mathcal{M}$ with L columns, representing the different amino-acid sites, and M rows. Each row i, denoted by $\boldsymbol{x}^{(i)}$, represents one sequence of the alignment. We refer to L as the MSA length, and to M as its depth. For all our MSAs, $M > 36000$. These alignments are the same as in Ref. [53], except that we discarded PF13354 (Beta-lactamase2) because of its smaller depth.

Deep MSAs generally include some highly similar sequences due to phylogenetic relatedness. This can be characterized via the effective depth [8]

$$M_{\text{eff}}^{(\delta)} := \sum_{i=1}^{M} w_i, \quad \text{with} \quad w_i := |\{i' : d_{\text{H}}(\boldsymbol{x}^{(i)}, \boldsymbol{x}^{(i')}) < \delta\}|^{-1}, \tag{8}$$

where $d_{\text{H}}(\boldsymbol{x}, \boldsymbol{y})$ is the (normalized) Hamming distance between two sequences $\boldsymbol{x}$ and $\boldsymbol{y}$, i.e. the fraction of sites where the amino acids differ, and we set $\delta = 0.2$. Note that the inverse of the sequence weight w_i in Eq. (8) is the number of neighbors in "Characterizing the distribution of sequences in MSAs", and that $M_{\text{eff}}^{(0.2)}/M$ can be as low as 0.06 for our natural MSAs.

All these families were previously shown to be well fitted by Potts models inferred by bmDCA [12], making our results on sequence generation by bmDCA readily comparable with previous results. Our domains' short lengths are convenient because bmDCA is computationally demanding, and also in view of MSA Transformer's large memory footprint, which is $O(LM^2) + O(L^2)$. Furthermore, their large depth is crucial to our comparisons, as it allows Potts model to be accurately fitted [12].

Pfam ID	Family name	L	M	$M_{\text{eff}}^{(0.2)}$
PF00004	AAA	132	39277	9049
PF00005	ABC_tran	137	68891	43881
PF00041	fn3	85	42721	17782
PF00072	Response_reg	112	73063	40180
PF00076	RRM_1	69	51964	20273
PF00096	zf-C2H2	23	38996	12581
PF00153	Mito_carr	94	93776	17859
PF00271	Helicase_C	111	66809	25017
PF00397	WW	31	39045	3361
PF00512	HisKA	66	154998	67303
PF00595	PDZ	82	71303	4053
PF01535	PPR	31	109064	37514
PF02518	HATPase_c	111	80714	59189
PF07679	I-set	90	36141	14611

Table 1: **Natural MSAs used in our analysis.** L denotes the length of an MSA, M its depth, and $M_{\text{eff}}^{(0.2)}$ its effective depth with distance threshold $\delta = 0.2$, see Eq. (8).

Results

An iterative masking procedure allows MSA Transformer to generate novel sequences with high scores

Is the protein language model MSA Transformer [25] able to generate sequences that are credible members of protein families? How do its generative abilities compare to Potts models inferred by Boltzmann machine DCA (bmDCA) [12], a state-of-the-art generative DCA method which has been experimentally shown to generate functional proteins [13]? To address these questions, we employed an iterative masking procedure to generate synthetic MSAs from natural MSAs of 14 different Pfam protein families (see Table 1) with MSA Transformer, as described in "Using MSA Transformer to generate sequences via an iterative masking procedure". We also generated synthetic sequences by MCMC sampling from Potts models inferred from these MSAs by bmDCA, see "Sampling sequences from Potts models". 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 (see "Scoring individual sequences"). 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. The three of them account for very different aspects of proteins, namely homology, coevolution and structure.

Fig. 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. Conversely, sequences generated by bmDCA feature scores that are on average worse than natural sequences. This

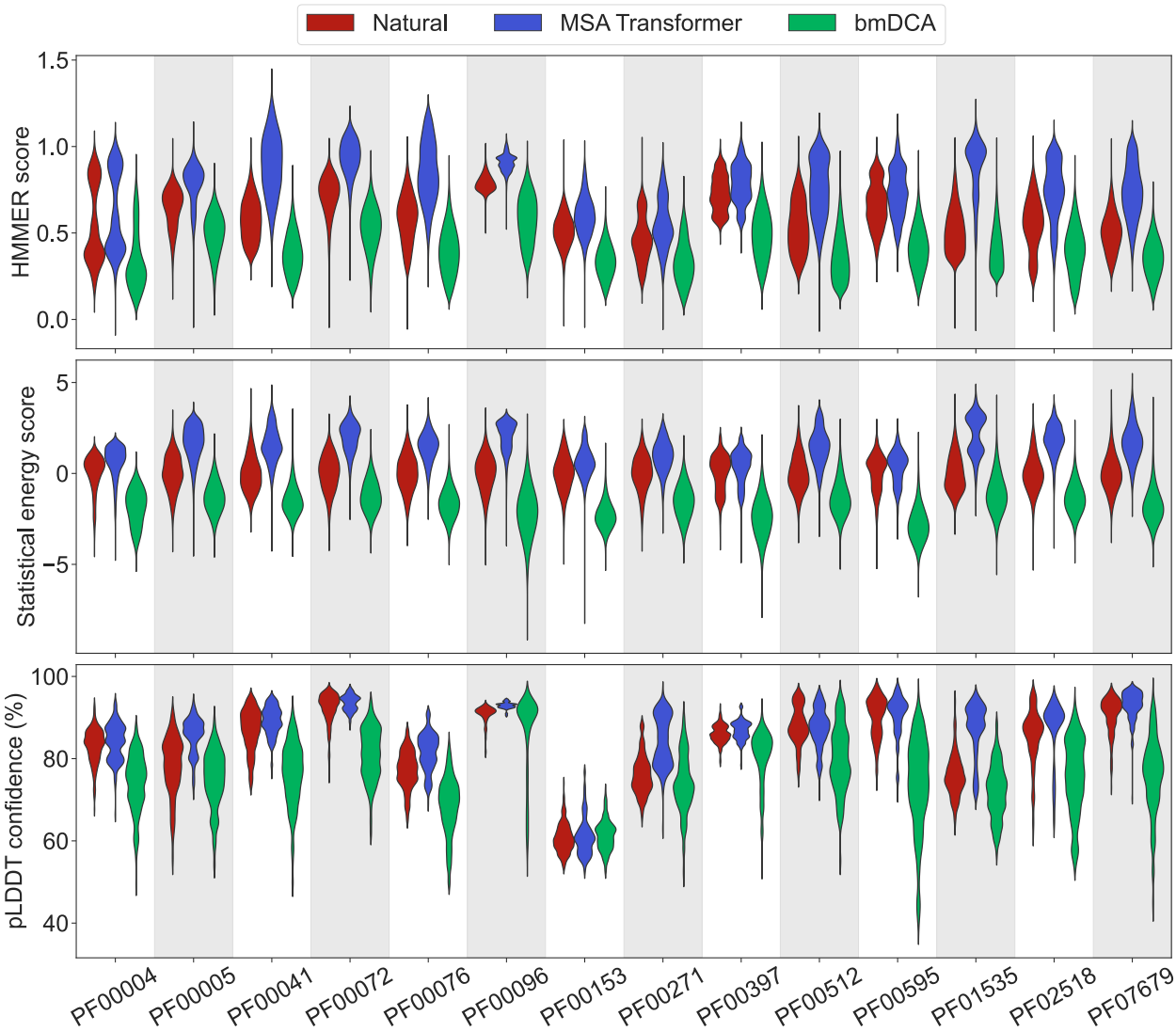

Figure 2: **Comparison of homology, coevolution, and structure-based scores between natural sequences and sequences generated by MSA Transformer or bmDCA.** For each Pfam family in Table 1, we compare a natural MSA from Pfam and two synthetic MSAs of the same depth. The first synthetic MSA was obtained using MSA Transformer via our iterative masking procedure, and the second one was generated by a Potts model inferred from the natural MSA using bmDCA. For each of the three scores described in "Scoring individual sequences", we show the distributions of score values among sequences in each MSA as a violin plot. Higher score values are better. **Top panel:** For each Pfam family, HMMER scores are divided by the highest score found in the natural MSA. Note that sequences below HMMER's default homology detection score (E-value larger than 10), and whose HMMER score is thus 0, are not shown (the median over families of the fraction of such sequences is 2% for bmDCA-generated MSAs, while there are no such sequences among the MSA Transformer-generated ones). **Middle panel:** Statistical energy scores are defined as minus the bmDCA statistical energies. To accommodate the highly family-dependent ranges of these scores, for each Pfam family we show their values after shifting by the mean score in the natural MSA, and normalizing by the standard deviation of natural MSA scores. **Bottom panel:** AlphaFold's pLDDT confidence scores coming from 200 randomly chosen sequences from each MSA. Note that all kernel-smoothed histograms are normalized such that all violins have the same maximal width.

is true even for a score built from the Potts model fitted by bmDCA, namely the statistical energy score. These results demonstrate that MSA Transformer is a good candidate to generate synthetic sequences from protein families, and that our iterative masking procedure allows to perform such generation.

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 Fig. 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 highly similar for the other families we studied.

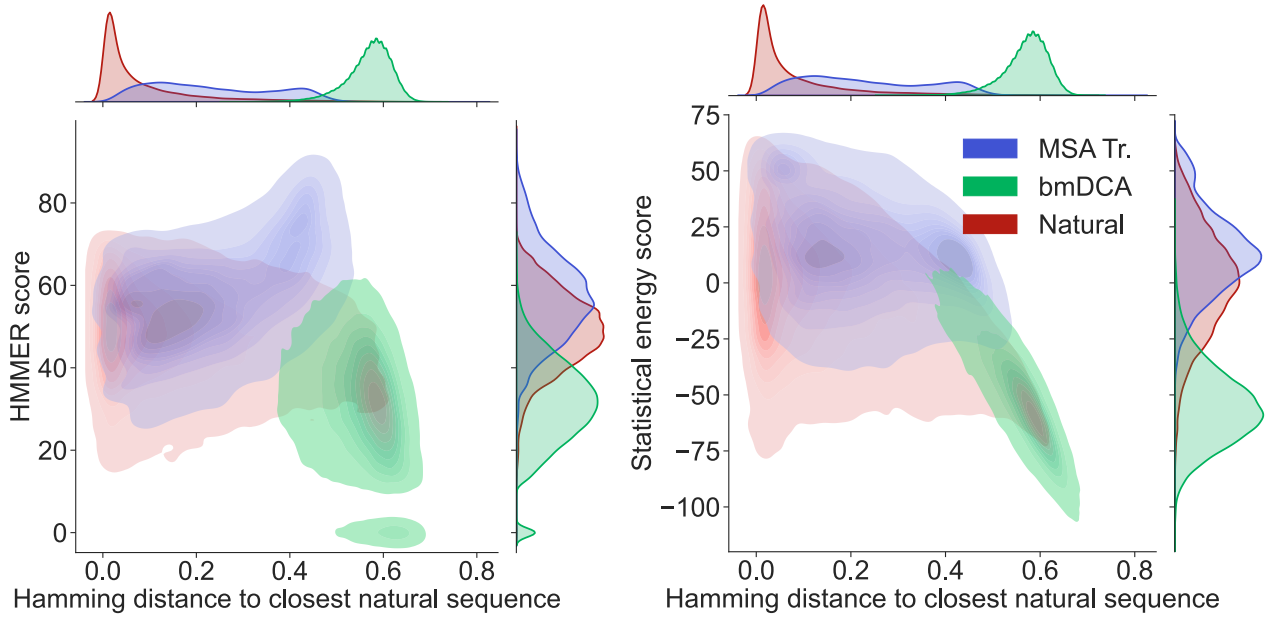

Figure 3: **Homology and coevolution scores vs. distance to the natural MSA, for protein family PF00153.** We show contour plots of the HMMER score and the statistical energy score (defined as minus the DCA statistical energy, shifted by its mean value in the natural MSA) versus the Hamming distance of each sequence to the closest natural sequence (which is not itself, in the case of natural sequences). Results are shown for natural sequences and for sequences generated using MSA Transformer and bmDCA. The lightest contours shown include 99% of the cumulative probability mass.

First, from the marginal distributions of the Hamming distances in Fig. 3, we observe 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. Second, the marginal distributions of scores illustrate the general observation made on Fig. 2 that MSA Transformer-generated sequences have higher overall scores than bmDCA generated ones and even than natural sequences. Moreover, the plots in Fig. 3 reveal that, 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. Quantitatively, 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. We observe these trends for most protein families studied. More precisely, MSA Transformer-generated MSAs yield

a median Pearson correlation of 0.31 with only 3 negative values out of 14, all satisfying $-0.2 < \rho < 0$. Conversely, all bmDCA-generated MSAs except one satisfy $\rho < -0.2$, and the median Pearson correlation is -0.52. In the case of the statistical energy score, the Pearson correlation values for PF00153 are $\rho = -0.30$ for MSA Transformer and $\rho = -0.87$ for bmDCA, which means that, for bmDCA, increasing Hamming distances to closest natural sequence leads to lower scores, while this anticorrelation is weaker for MSA Transformer. Across our 14 families, the median Pearson correlation is -0.31 for MSA Transformer and -0.92 for bmDCA, showing that the results obtained for PF00153 are typical. While our analysis was conducted using Hamming distances to assess similarity between sequences, we checked that the same conclusions are obtained when using BLOSUM similarity scores instead (from https://www.ncbi.nlm.nih.gov/Class/FieldGuide/BLOSUM62.txt). 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 (see "Sampling sequences from Potts models") was proposed as a mitigating strategy [13]. However, here, we observed that this substantially decreased the fitting of the first- and second-order statistics, and only slightly improved HMMER scores.

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, we compute one-, two- and three-body frequencies from our natural and synthetic MSAs and compared them for these statistics and their information-theory generalizations (see "Analyzing the statistics of MSAs"). Fig. 4 shows a comparison of second-order connected correlations for PF00153. We find that MSA Transformer reproduces these correlations less accurately than bmDCA. 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, while MSA Transformer has entirely different training objectives. Thus, this result is in line with expectations. What about higher-order statistics which are not fitting objectives of bmDCA? In the second row of Fig. 4, we show a comparison of third-order connected correlations, and in Fig. S3, we show histograms of co-information in natural and synthetic histograms, still for PF00153. These figures show that 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, as shown in Fig. S4.

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 [53]. Besides, inferred Potts models are known to be impacted by phylogenetic correlations [8, 10, 65–71].

First, to assess whether generated sequences most resemble natural ones that are well represented in their family or, rather, rare ones, we consider the closest natural sequence to each synthetic sequence, and count the neighbors of this natural sequence in the natural MSA (see "Characterizing the distribution of sequences in MSAs"). Fig. 5 compares the distribution of these numbers of neighbors for natural sequences and for the closest natural sequences to gen-

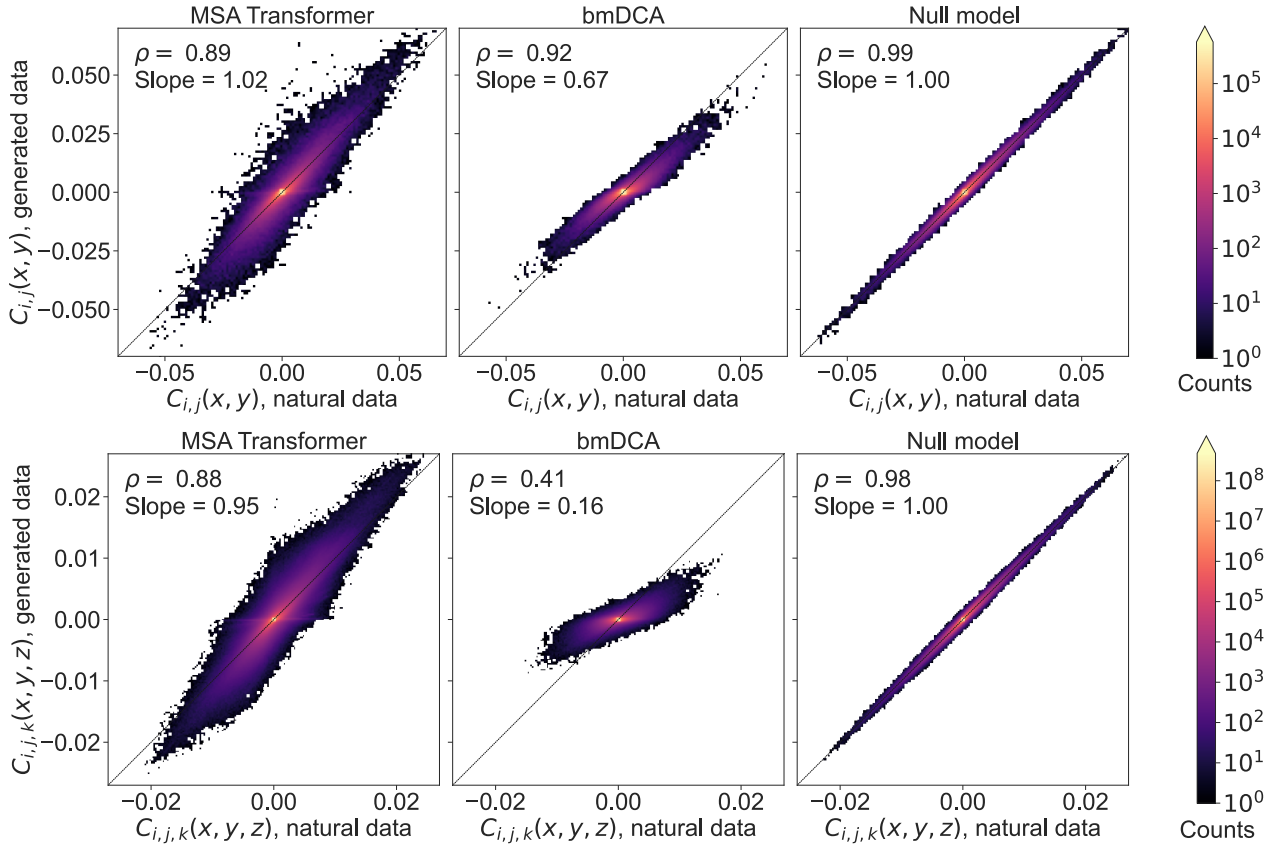

Figure 4: **Two- and three-body connected correlations estimated from generated MSAs versus the natural one, for family PF00153.** Relationships between connected correlations estimated from the MSA generated by MSA Transformer or bmDCA, and those estimated from the natural MSA, are shown as binned scatter 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. For the comparisons involving MSA Transformer, we used Eqs. (4) and (5) to estimate correlations. On the other hand, since bmDCA is trained to reproduce frequencies rescaled with phylogenetic weights (w_i in Eq. (8)), for the comparisons involving bmDCA we rescaled the natural frequencies before using Eqs. (4) and (5) to estimate correlations.

erated sequences, in the case of PF00153. It shows that bmDCA generates sequences similar to natural sequences with few neighbors. Conversely, MSA Transformer generates sequences whose closest natural sequences have a distribution of number of neighbors similar to that of the natural MSA, and it produces slightly fewer sequences with few neighbors than there are in the natural MSA.

To analyze the distribution of MSA sequences in sequence space in more detail, we perform a principal component analysis of one-hot encoded MSAs, and focus on the top two principal components (PCs) [12] (see "Characterizing the distribution of sequences in MSAs"). Fig. 6 shows the distribution of sequences in the space spanned by these top two PCs, for natural and synthetic 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 that of the natural MSA. Conversely, bmDCA captures the overall shape of this distribution, but appears to smooth it compared to the natural data. This observation is general across all the MSAs we considered.

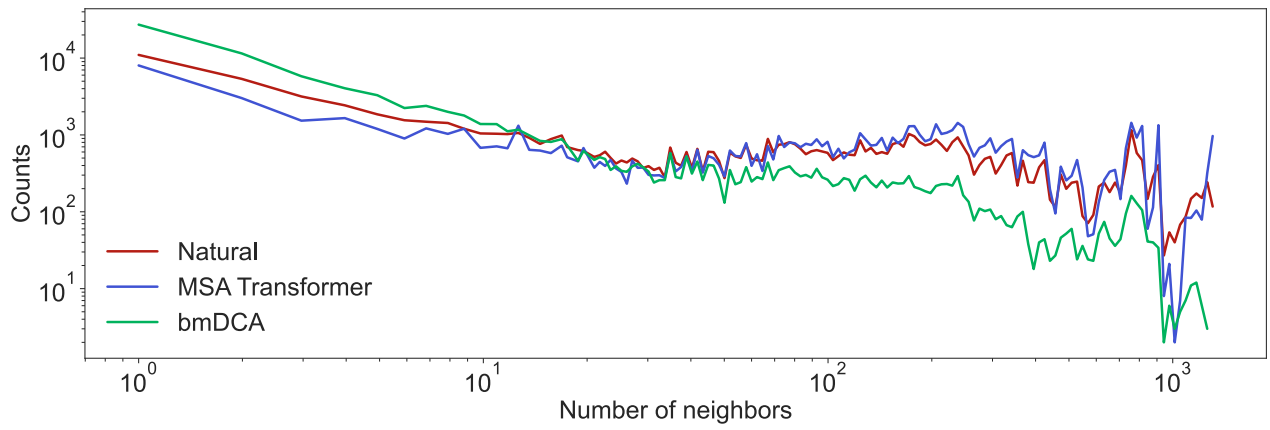

Figure 5: **Neighbors of natural and synthetic sequences, for family PF00153.** We show the distribution of the number of neighbors of sequences in the natural MSA, and the distribution of the number of neighbors of the closest natural sequence to each of our generated sequences. Given a sequence in a natural MSA, its number of neighbors is the number of natural sequences that are within a (normalized) Hamming distance $\delta = 0.2$ from it.

Note that a limitation of this analysis is that the top two PCs explain a small fraction of the variance in all cases (see Fig. 6).

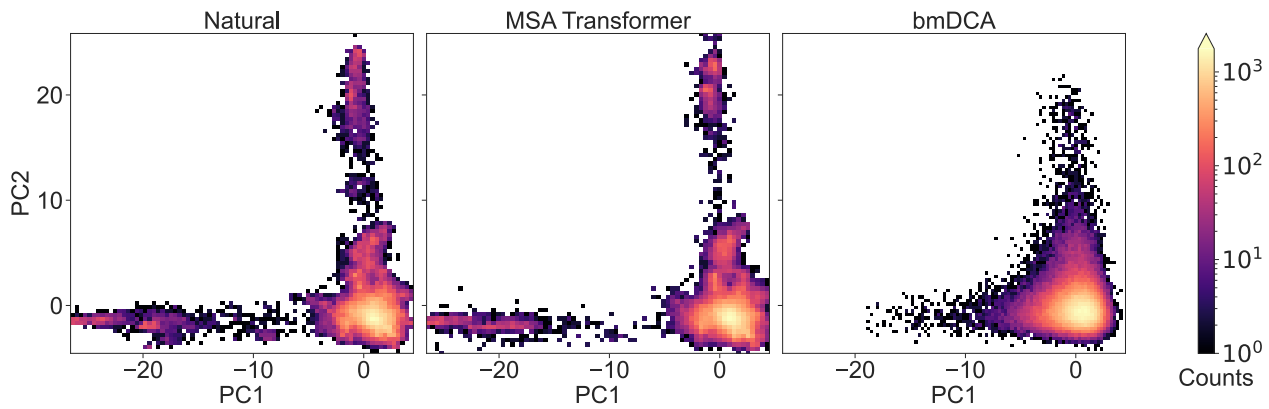

Figure 6: **Distribution of sequences in sequence space, for family PF00153.** We show the distribution of one-hot encoded natural and synthetic sequences projected in the subspace of the first two principal components of the natural MSA. Note that the variances explained by the first two principal components of each MSA are respectively: Natural, 1.8%; MSA Transformer, 2.9%; bmDCA, 0.8%.

Finally, to analyze in more detail the apparent relatedness of generated sequences, and compare it to real phylogenetic relationships in natural sequences, we infer phylogenetic trees from each synthetic and natural MSA, and analyze the eigenvalue spectrum of their modified graph Laplacian (MGL) to compare them [62] (see "Characterizing the distribution of sequences in MSAs"). Fig. S5 compares the density of these eigenvalue spectra for natural and synthetic MSAs regarding PF00153, considering all tree leaves. The skewness and the position of such distributions are indicators of the topology of the tree. In particular, distributions with negative skewness (right unbalanced) or which are shifted more to the right, correspond to "tippy" trees, while the opposite case corresponds to "stemmy" trees [62], which feature an accumulation of recent speciation events (short leaves length) [72]. In this light, Fig. S5 shows that, focusing on tree leaves, both MSA Transformer and bmDCA generate sequences with an apparent phylogeny that differs substantially from the natural one. Furthermore, while bmDCA-generated

sequences give trees which are more tippy than the natural, MSA Transformer gives more stemmy ones, reflecting the fact that bmDCA sequences are more distant from one another.

Discussion

In this work, we proposed an iterative masking procedure which directly exploits the masked language modeling objective of protein language models 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. Conversely, bmDCA better reproduces first- and second-order statistics, consistently with its training objective.

Our results are highly promising for sequence generation by MSA-based protein language models, and we hope that they will motivate further studies, especially experimental tests. More generally, our results reinforce the new promising "coevolution-driven" protein design approach of learning from sequences of evolutionarily related proteins the constraints associated to protein structure and function. This concept differs from structure- and physics-based *de novo* design [3–5], and from the new possibility to use supervised deep learning models able to accurately predict protein structures [28, 73, 74] for structure-driven sequence generation [75]. One can view the coevolution-driven approach as intermediate between structure-based approaches and directed evolution ones [7]. The coevolution-driven approach was recently experimentally validated in the case of bmDCA Potts models, which capture pairwise coevolution patterns in MSAs [13], and for variational autoencoders [14, 15]. Protein language models trained on multiple sequence alignments provide state-of-the-art unsupervised contact prediction and are able to capture coevolutionary patterns in their tied row attentions [25], and capture phylogenetic relationships in column attentions [53]. This makes them ideal candidates to generate new protein sequences from given families. However, contrary to Potts models and variational autoencoders [15], they do not allow direct sampling from a probability distribution over sequences [50]. Here, we demonstrated the power of a simple generation method directly based on the masked language modeling objective used for the training of MSA-based protein language models. It differs from using a decoder in this context, which allows autoregressive generation of amino acids to form a new sequence, but requires training a full encoder-decoder model and learning a parametric function mapping an MSA to a distribution over its sequences [30]. 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 along this line [50], but our minimal approach already gives very promising results.

We have focused on a large protein language model and compared it to the simplest model capturing coevolution, namely the Potts model, but we note that interpretable models of intermediate complexity such as restricted Boltzmann machines [76] could also be explored for coevolution-driven protein design. All these methods rely on MSAs; this is very useful to capture coevolution, but also means that one has to rely on potentially imperfect alignments. Thus, starting from alignment-free methods [24, 77, 78] also constitutes a promising direction.

Acknowledgments

This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement No. 851173, to A.-F. B.).

References

[1] Michael Socolich, Steve W. Lockless, William P. Russ, Heather Lee, Kevin H. Gardner, and Rama Ranganathan. Evolutionary information for specifying a protein fold. *Nature*, 437(7058):512–518, 2005.

[2] William Bialek. *Biophysics: Searching for Principles*. Princeton University Press, 2012.

[3] B. I. Dahiyat and S. L. Mayo. De novo protein design: fully automated sequence selection. *Science*, 278(5335):82–87, Oct 1997.

[4] B. Kuhlman, G. Dantas, G. C. Ireton, G. Varani, B. L. Stoddard, and D. Baker. Design of a novel globular protein fold with atomic-level accuracy. *Science*, 302(5649):1364–1368, Nov 2003.

[5] Huanhuan Liang, Hao Chen, Keqiang Fan, Ping Wei, Xianrong Guo, Changwen Jin, Chen Zeng, Chao Tang, and Luhua Lai. De novo design of a $\beta\alpha\beta$ motif. *Angewandte Chemie International Edition*, 48(18):3301–3303, 2009.

[6] G. J. Rocklin, T. M. Chidyausiku, I. Goreshnik, A. Ford, S. Houliston, A. Lemak, L. Carter, R. Ravichandran, V. K. Mulligan, A. Chevalier, C. H. Arrowsmith, and D. Baker. Global analysis of protein folding using massively parallel design, synthesis, and testing. *Science*, 357(6347):168–175, 07 2017.

[7] Frances H. Arnold. Directed evolution: Bringing new chemistry to life. *Angewandte Chemie International Edition*, 57(16):4106–4106, 2018.

[8] 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. U.S.A.*, 106(1):67–72, Jan 2009.

[9] Faruck Morcos, Andrea Pagnani, Bryan Lunt, Arianna Bertolino, Debora S. Marks, Chris Sander, Riccardo Zecchina, José N. Onuchic, Terence Hwa, and Martin Weigt. Direct-coupling analysis of residue coevolution captures native contacts across many protein families. *Proceedings of the National Academy of Sciences of the United States of America*, 108(49), 2011.

[10] 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(12):e28766, 2011.

[11] Magnus Ekeberg, Cecilia Lövkvist, Yueheng Lan, Martin Weigt, and Erik Aurell. Improved contact prediction in proteins: Using pseudolikelihoods to infer Potts models. *Physical Review E - Statistical, Nonlinear, and Soft Matter Physics*, 87(1):1–16, 2013.

[12] Matteo Figliuzzi, Pierre Barrat-Charlaix, and Martin Weigt. How pairwise coevolutionary models capture the collective residue variability in proteins? *Molecular Biology and Evolution*, 35(4):1018–1027, 2018.

[13] William P. Russ, Matteo Figliuzzi, Christian Stocker, Pierre Barrat-Charlaix, Michael Socolich, Peter Kast, Donald Hilvert, Remi Monasson, Simona Cocco, Martin Weigt, and Rama Ranganathan. An evolution-based model for designing chorismate mutase enzymes. *Science*, 369(6502):440–445, 2020.

[14] 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(2):e1008736, 02 2021.

[15] Francisco McGee, Sandro Hauri, Quentin Novinger, Slobodan Vucetic, Ronald M. Levy, Vincenzo Carnevale, and Allan Haldane. The generative capacity of probabilistic protein sequence models. *Nat. Commun.*, 12(1):6302, 11 2021.

[16] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural Machine Translation by Jointly Learning to Align and Translate (ICLR 2015), 2014.

[17] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in Neural Information Processing Systems*, 30:5998–6008, 2017.

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

[19] Ahmed Elnaggar, Michael Heinzinger, Christian Dallago, Ghalia Rehawi, Yu Wang, Llion Jones, Tom Gibbs, Tamas Feher, Christoph Angerer, Martin Steinegger, Debsindhu Bhowmik, and Burkhard Rost. ProtTrans: Towards cracking the language of life's code through self-supervised learning, 2020.

[20] Alexander Rives, Joshua Meier, Tom Sercu, Siddharth Goyal, Zeming Lin, Jason Liu, Demi Guo, Myle Ott, C. Lawrence Zitnick, Jerry Ma, and Rob Fergus. Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *Proc. Natl. Acad. Sci. U.S.A.*, 118(15), 2021.

[21] Roshan Rao, Joshua Meier, Tom Sercu, Sergey Ovchinnikov, and Alexander Rives. Transformer protein language models are unsupervised structure learners. In *International Conference on Learning Representations*, 2021.

[22] Jesse Vig, Ali Madani, Lav R. Varshney, Richard Socher, Caiming Xiong, and Nazneen Fatema Rajani. BERTology meets Biology: Interpreting attention in Protein language Models. 2021.

[23] Ali Madani, Bryan McCann, Nikhil Naik, Nitish Shirish Keskar, Namrata Anand, Raphael R. Eguchi, Po-Ssu Huang, and Richard Socher. ProGen: Language Modeling for Protein Generation. 2020.

[24] Ali Madani, Ben Krause, Eric R. Greene, Subu Subramanian, Benjamin P. Mohr, James M. Holton, Jose Luis Olmos, Caiming Xiong, Zachary Z. Sun, Richard Socher, James S. Fraser, and Nikhil Naik. Deep neural language modeling enables functional protein generation across families, 2021.

[25] Roshan M Rao, Jason Liu, Robert Verkuil, Joshua Meier, John Canny, Pieter Abbeel, Tom Sercu, and Alexander Rives. MSA transformer. 139:8844–8856, 18–24 Jul 2021.

[26] Nicholas Bhattacharya, Neil Thomas, Roshan Rao, Justas Dauparas, Peter K. Koo, David Baker, Yun S. Song, and Sergey Ovchinnikov. Single layers of attention suffice to predict protein contacts. *bioRxiv*, 2020.

[27] Nicholas Bhattacharya, Neil Thomas, Roshan Rao, Justas Dauparas, Peter K Koo, David Baker, Yun S Song, and Sergey Ovchinnikov. Interpreting Potts and Transformer Protein Models Through the Lens of Simplified Attention. *arXiv*, pages 34–45, 2021.

[28] John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, Alex Bridgland, Clemens Meyer, Simon A A Kohl, Andrew J Ballard, Andrew Cowie, Bernardino Romera-Paredes, Stanislav Nikolov, Rishub Jain, Jonas Adler, Trevor Back, Stig Petersen, David Reiman, Ellen Clancy, Michal Zielinski, Martin Steinegger, Michalina Pacholska, Tamas Berghammer, Sebastian Bodenstein, David Silver, Oriol Vinyals, Andrew W Senior, Koray Kavukcuoglu, Pushmeet Kohli, and Demis Hassabis. Highly accurate protein structure prediction with AlphaFold. *Nature*, 2021.

[29] Sean R Johnson, Kenneth Massie, Sarah Monaco, and Zaid Syed. Generating novel protein sequences using Gibbs sampling of masked language models. *bioRxiv*, page 2021.01.26.428322, 2021.

[30] Alex Hawkins-Hooker, David T. Jones, and Brooks Paige. MSA-conditioned generative protein language models for fitness landscape modelling and design. In *Machine Learning for Structural Biology Workshop, NeurIPS*, 2021.

[31] Noelia Ferruz, Steffen Schmidt, and Birte Höcker. A deep unsupervised language model for protein design. *bioRxiv*, 2022.

[32] Brian L Hie, Duo Xu, Varun R Shanker, Theodora UJ Bruun, Payton A Weidenbacher, Shaogeng Tang, and Peter S Kim. Efficient evolution of human antibodies from general protein language models and sequence information alone. *bioRxiv*, 2022.

[33] Georg Casari, Chris Sander, and Alfonso Valencia. A method to predict functional residues in proteins. *Nature*, 2(2):171–178, 1995.

[34] A. S. Lapedes, B. G. Giraud, L. Liu, and G. D. Stormo. Correlated mutations in models of protein sequences: phylogenetic and structural effects. In *Statistics in molecular biology and genetics - IMS Lecture Notes - Monograph Series*, volume 33, pages 236–256. 1999.

[35] S. D. Dunn, L. M. Wahl, and G. B. Gloor. Mutual information without the influence of phylogeny or entropy dramatically improves residue contact prediction. *Bioinformatics*, 24(3):333–340, Feb 2008.

[36] W. Bialek and R. Ranganathan. Rediscovering the power of pairwise interactions. *ArXiv*, 2007.

[37] J. I. Sułkowska, F. Morcos, M. Weigt, T. Hwa, and J. N. Onuchic. Genomics-aided structure prediction. *Proc. Natl. Acad. Sci. U.S.A.*, 109(26):10340–10345, Jun 2012.

[38] R. S. Dwyer, D. P. Ricci, L. J. Colwell, T. J. Silhavy, and N. S. Wingreen. Predicting functionally informative mutations in Escherichia coli BamA using evolutionary covariance analysis. *Genetics*, 195(2):443–455, Oct 2013.

[39] R. R. Cheng, F. Morcos, H. Levine, and J. N. Onuchic. Toward rationally redesigning bacterial two-component signaling systems using coevolutionary information. *Proc. Natl. Acad. Sci. U.S.A.*, 111(5):E563–571, Feb 2014.

[40] R. R. Cheng, O. Nordesjo, R. L. Hayes, H. Levine, S. C. Flores, J. N. Onuchic, and F. Morcos. Connecting the Sequence-Space of Bacterial Signaling Proteins to Phenotypes Using Coevolutionary Landscapes. *Mol. Biol. Evol.*, 33(12):3054–3064, 12 2016.

[41] M. Figliuzzi, H. Jacquier, A. Schug, O. Tenaillon, and M. Weigt. Coevolutionary Landscape Inference and the Context-Dependence of Mutations in Beta-Lactamase TEM-1. *Mol. Biol. Evol.*, 33(1):268–280, Jan 2016.

[42] J. A. de la Paz, C. M. Nartey, M. Yuvaraj, and F. Morcos. Epistatic contributions promote the unification of incompatible models of neutral molecular evolution. *Proc Natl Acad Sci U S A*, 117(11):5873–5882, 03 2020.

[43] F. Morcos, B. Jana, T. Hwa, and J. N. Onuchic. Coevolutionary signals across protein lineages help capture multiple protein conformations. *Proc. Natl. Acad. Sci. U.S.A.*, 110(51):20533–20538, Dec 2013.

[44] D. Malinverni, S. Marsili, A. Barducci, and P. De Los Rios. Large-Scale Conformational Transitions and Dimerization Are Encoded in the Amino-Acid Sequences of Hsp70 Chaperones. *PLoS Comput. Biol.*, 11(6):e1004262, Jun 2015.

[45] Anne Florence Bitbol, Robert S. Dwyer, Lucy J. Colwell, and Ned S. Wingreen. Inferring interaction partners from protein sequences. *Proceedings of the National Academy of Sciences of the United States of America*, 113(43):12180–12185, 2016.

[46] T. Gueudre, C. Baldassi, M. Zamparo, M. Weigt, and A. Pagnani. Simultaneous identification of specifically interacting paralogs and interprotein contacts by direct coupling analysis. *Proc. Natl. Acad. Sci. U.S.A.*, 113(43):12186–12191, 10 2016.

[47] Qian Cong, Ivan Anishchenko, Sergey Ovchinnikov, and David Baker. Protein interaction networks revealed by proteome coevolution. *Science*, 365(6449):185–189, 2019.

[48] Anna G. Green, Hadeer Elhabashy, Kelly P. Brock, Rohan Maddamsetti, Oliver Kohlbacher, and Debora S. Marks. Large-scale discovery of protein interactions at residue resolution using co-evolution calculated from genomic sequences. *Nature Communications*, 12(1):1–12, 2021.

[49] Alex Wang and Kyunghyun Cho. BERT has a mouth, and it must speak: BERT as a Markov random field language model. *CoRR*, abs/1902.04094, 2019.

[50] Kartik Goyal, Chris Dyer, and Taylor Berg-Kirkpatrick. Exposing the implicit energy networks behind masked language models via Metropolis–Hastings, 2021.

[51] Sean R. Eddy. Profile hidden Markov models. *Bioinformatics*, 14(9):755–763, 1998.

[52] Joshua Meier, Roshan Rao, Robert Verkuil, Jason Liu, Tom Sercu, and Alexander Rives. Language models enable zero-shot prediction of the effects of mutations on protein function. pages 1–28, 2021.

[53] Umberto Lupo, Damiano Sgarbossa, and Anne-Florence Bitbol. Protein language models trained on multiple sequence alignments learn phylogenetic relationships, 2022.

[54] S. Cocco, C. Feinauer, M. Figliuzzi, R. Monasson, and M. Weigt. Inverse statistical physics of protein sequences: a key issues review. *Rep Prog Phys*, 81(3):032601, 03 2018.

[55] William J. McGill. Multivariate information transmission. *Psychometrika*, 19(2):97–116, jun 1954.

[56] N. Timme, W. Alford, B. Flecker, and J. M. Beggs. Synergy, redundancy, and multivariate information measures: an experimentalist's perspective. *J Comput Neurosci*, 36(2):119–140, Apr 2014.

[57] R. Quax, O. Har-Shemesh, and P. M. A. Sloot. Quantifying Synergistic Information Using Intermediate Stochastic Variables. *Entropy*, 19:85, 2017.

[58] F. E. Rosas, P. A. M. Mediano, M. Gastpar, and H. J. Jensen. Quantifying high-order interdependencies via multivariate extensions of the mutual information. *Phys Rev E*, 100(3-1):032305, Sep 2019.

[59] Paul L. Williams and Randall D. Beer. Nonnegative decomposition of multivariate information, 2010.

[60] Fernando Rosas, Vasilis Ntranos, Christopher J. Ellison, Sofie Pollin, and Marian Verhelst. Understanding interdependency through complex information sharing. *Entropy*, 18(2), 2016.

[61] Morgan N. Price, Paramvir S. Dehal, and Adam P. Arkin. FastTree 2 – Approximately Maximum-Likelihood Trees for Large Alignments. *PLoS ONE*, 5(3):e9490, mar 2010.

[62] Eric Lewitus and Helene Morlon. Characterizing and comparing phylogenies from their laplacian spectrum. *Systematic Biology*, 65(3):495–507, 2016.

[63] C. Colijn and G. Plazzotta. A Metric on Phylogenetic Tree Shapes. *Systematic Biology*, 67(1):113–126, 2018.

[64] Jaina Mistry, Sara Chuguransky, Lowri Williams, Matloob Qureshi, Gustavo A Salazar, Erik L L Sonnhammer, Silvio C E Tosatto, Lisanna Paladin, Shriya Raj, Lorna J Richardson, Robert D Finn, and Alex Bateman. Pfam: The protein families database in 2021. *Nucleic Acids Research*, 49(D1):D412–D419, 10 2020.

[65] C. Qin and L. J. Colwell. Power law tails in phylogenetic systems. *Proc. Natl. Acad. Sci. U.S.A.*, 115(4):690–695, Jan 2018.

[66] Susann Vorberg, Stefan Seemayer, and Johannes Söding. Synthetic protein alignments by ccmgen quantify noise in residue-residue contact prediction. *PLOS Computational Biology*, 14(11):1–25, 11 2018.

[67] Edwin Rodriguez Horta, Pierre Barrat-Charlaix, and Martin Weigt. Toward inferring potts models for phylogenetically correlated sequence data. *Entropy*, 21(11), 2019.

[68] Edwin Rodriguez Horta and Martin Weigt. On the effect of phylogenetic correlations in coevolution-based contact prediction in proteins. *PLoS Comput Biol*, 17(5), 2021.

[69] G. Marmier, M. Weigt, and A.-F. Bitbol. Phylogenetic correlations can suffice to infer protein partners from sequences. *PLoS Comput. Biol.*, 15(10):e1007179, Oct 2019.

[70] A. Colavin, E. Atolia, A.-F. Bitbol, and K. C. Huang. Extracting phylogenetic dimensions of coevolution reveals hidden functional signals. *Scientific Reports*, 12:820, 2022.

[71] Andonis Gerardos, Nicola Dietler, and Anne-Florence Bitbol. Correlations from structure and phylogeny combine constructively in the inference of protein partners from sequences. *bioRxiv*, 2021.

[72] Rafael Molina-Venegas. What are "tippy" and "stemmy" phylogenies? Resolving a phylogenetic terminological tangle. *Journal of Systematics and Evolution*, 59(2):403–404, mar 2021.

[73] Minkyung Baek, Frank Dimaio, Ivan Anishchenko, and Justas Dauparas. Accurate prediction of protein structures and interactions using a 3-track network. pages 1–17, 2021.

[74] Ratul Chowdhury, Nazim Bouatta, Surojit Biswas, Charlotte Rochereau, George M. Church, Peter K. Sorger, and Mohammed AlQuraishi. Single-sequence protein structure prediction using language models from deep learning. *bioRxiv*, 2021.

[75] I. Anishchenko, S. J. Pellock, T. M. Chidyausiku, T. A. Ramelot, S. Ovchinnikov, J. Hao, K. Bafna, C. Norn, A. Kang, A. K. Bera, F. DiMaio, L. Carter, C. M. Chow, G. T. Montelione, and D. Baker. De novo protein design by deep network hallucination. *Nature*, Dec 2021.

[76] Jérôme Tubiana, Simona Cocco, and Rémi Monasson. Learning protein constitutive motifs from sequence data. *eLife*, 8, 2019.

[77] Maxwell L. Bileschi, David Belanger, Drew Bryant, Theo Sanderson, Brandon Carter, D. Sculley, Mark A. DePristo, and Lucy J. Colwell. Using Deep Learning to Annotate the Protein Universe. *bioRxiv*, pages 1–29, 2019.

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

Supplementary material

	Natural sequences	MSA Transformer					bmDCA		
		Iterative masking			Context (greedy)		MCMC sampling		
Score		Greedy	$T = 0.5$	$T = 1.0$	Fixed	Variable	$T = 0.9$	$T = 0.95$	$T = 1$
Distance	0.155	0.271	0.305	0.514	0.232	0.362	0.432	0.527	0.585
HMMER	48.0	58.2	58.1	48.4	58.7	63.8	37.2	35.1	31.4
−Energy	0	13.0	8.5	−42.0	−15.4	−13.2	−2.1	−32.9	−54.7
$\rho\left[C_{ij}\right]$	0.94	0.84	0.84	0.62	0.73	0.81	0.50	0.68	0.78
$\rho\left[C_{ijk}\right]$	0.89	0.80	0.76	0.41	0.66	0.77	0.30	0.36	0.20

Table S1: **Comparing different generation methods.** Various scores are shown for the natural MSA of protein family PF00153 and for synthetic MSAs generated in different ways from this family (each synthetic MSA comprises 10,000 sequences). For generation using MSA Transformer (see "Using MSA Transformer to generate sequences via an iterative masking procedure"), our standard iterative masking procedure is shown with its default greedy sampling (corresponding to $T = 0$) and two higher temperatures. Variants of the procedure where only the first sequence is masked ("Context", either fixed or variable, both with greedy sampling) are also shown. For generation using bmDCA, three temperature values are shown, the default being $T = 1$. We report the mean Hamming distance to the closest natural sequence ("Distance") as well as the mean HMMER and statistical energy scores ("−Energy") described in "Scoring individual sequences". Note that statistical energy scores are shifted by the mean value obtained for the natural MSA (which is −235.8). We also report the Pearson correlations between the two- and three-body statistics of the natural and the generated MSAs, denoted respectively by $\rho\left[C_{ij}\right]$ and $\rho\left[C_{ijk}\right]$ (for the natural MSA we report the Pearson correlation between two halves of this MSA), as illustrated in Fig. 4.

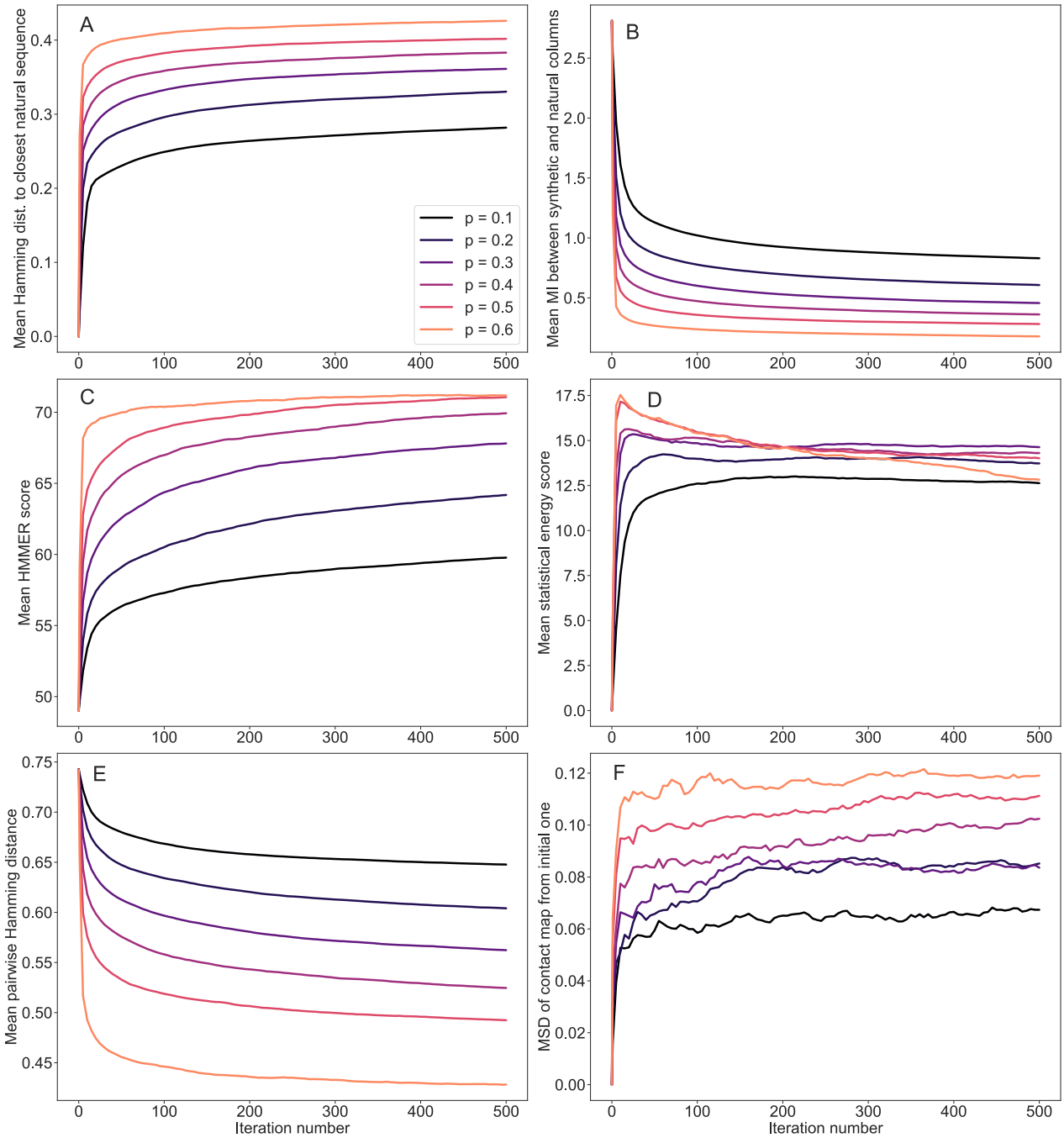

Figure S1: **Evolution of mean scores during the iterative masking procedure, for family PF00153.** Average scores of the generated sequences are reported for different iteration numbers and masking probabilities. The scores employed are: (**A**) Hamming distances to the closest natural sequence, (**B**) Mutual information between synthetic and natural columns of the MSAs, (**C**) HMMER scores – see "Scoring individual sequences", (**D**) Statistical energy scores (negative DCA statistical energies, shifted by their mean value for natural sequences – see "Scoring individual sequences"), (**E**) Pairwise Hamming distances between the generated sequences, (**F**) Mean square deviations (MSD) of the predicted contact maps at each iteration from the initial one (zero iterations). The synthetic MSAs, comprising 5000 sequences, were generated with MSA Transformer starting from natural sequences of family PF00153.

Figure S2: **Evolution of inferred contact maps during the iterative masking procedure, for family PF00153.** Contact maps obtained from MSA Transformer for different iteration numbers and masking probabilities are reported. Probabilities of contacts are computed using the logistic regression on the output attention matrices from Ref. [25]. The input of the model consists of 100 different sequences chosen uniformly at random from the synthetic MSA generated at each iteration of our iterative masking procedure. These contact maps were employed to compute the mean square deviations shown in Fig. S1**F**.

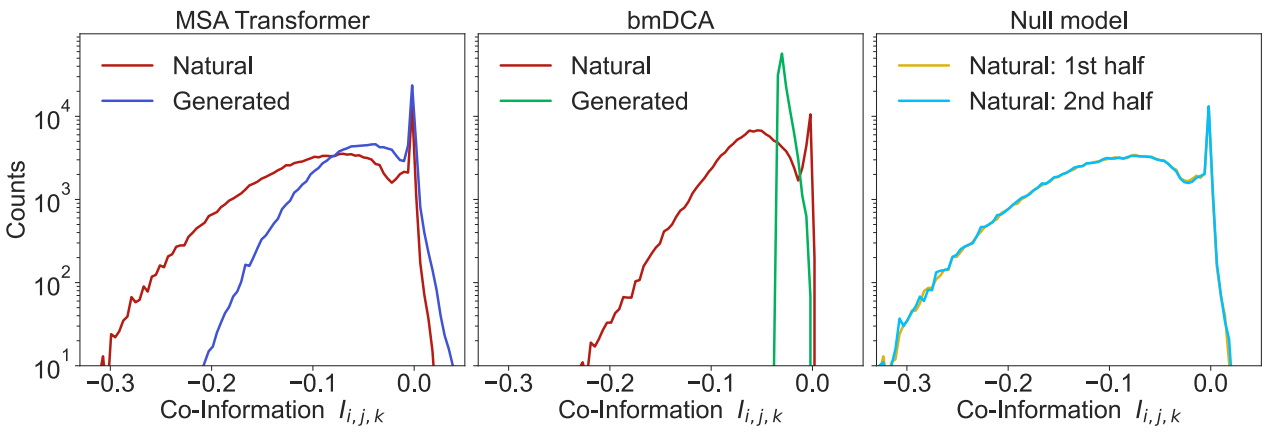

Figure S3: **Co-information histograms for all triplets of columns in the natural MSA and in the generated MSAs, for family PF00153.** For all triplets of sites, we estimate the co-information between the corresponding columns. For the comparisons involving MSA Transformer, we use Eq. (7) to estimate co-informations. On the other hand, since bmDCA is trained to reproduce frequencies rescaled with phylogenetic weights, for the comparisons involving bmDCA we rescale the natural frequencies using Eq. (7) to estimate co-informations. To assess fluctuations arising from finite-size effects, we include a null model (third column) obtained by splitting the natural MSA in half and comparing the co-informations of one half with those of the other.

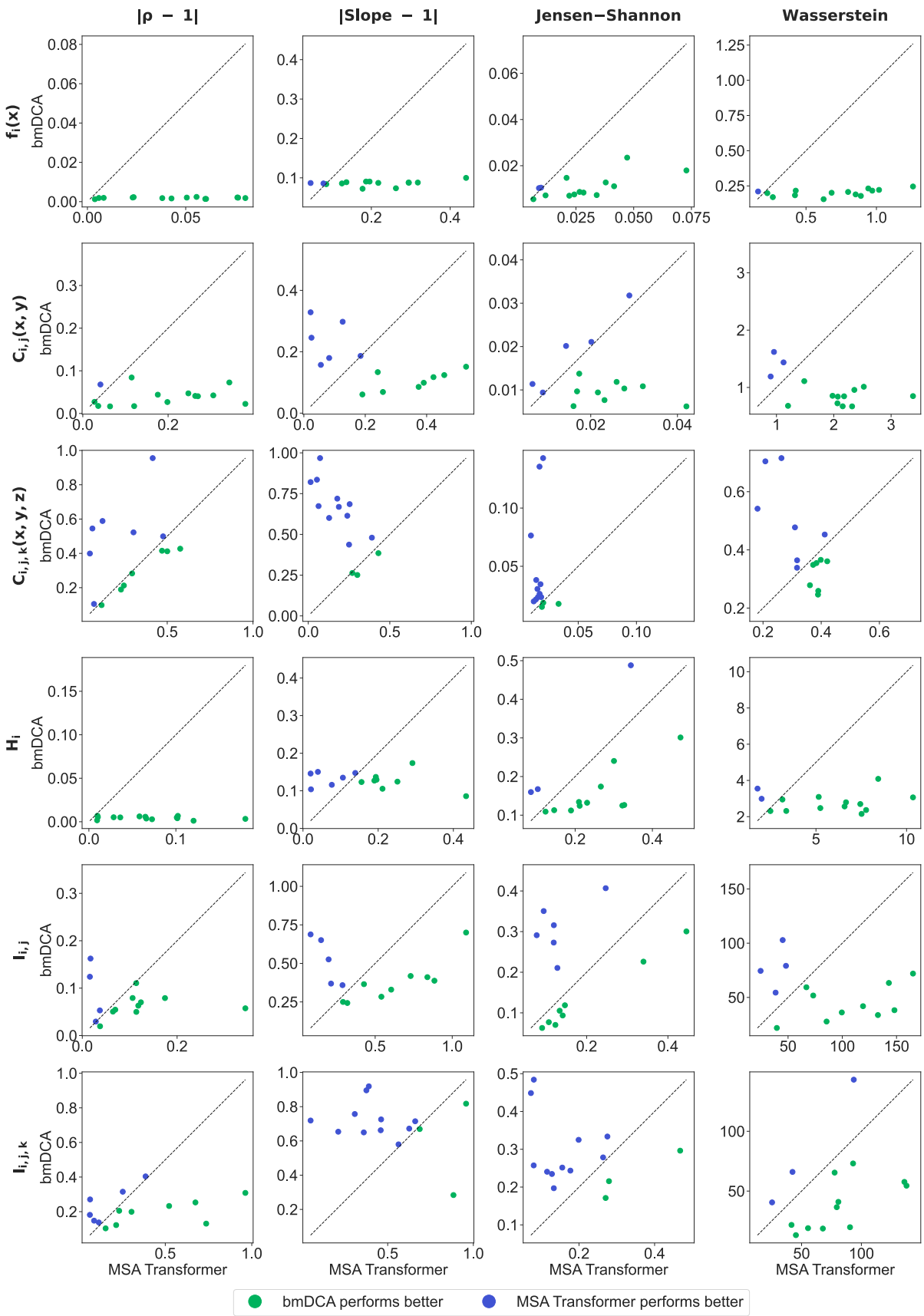

Figure S4: **Ability of generated sequences to reproduce one-, two- and three-body statistics (details on next page).** In all panels, each marker represents a family in Table 1. For each statistical or information measure (rows), different scores (columns) comparing generated and natural sequences are the coordinates of these points. All scores are such that being close to 0 is better, and their value for MSAs generated by bmDCA is shown versus that for MSAs generated by MSA Transformer.

Figure S4: **Ability of generated sequences to reproduce one-, two- and three-body statistics (continued).** The statistical or information measures considered in each row are defined in "Analyzing the statistics of MSAs" – from the top: one-body frequency, two- and three-body connected correlations, entropy, mutual information and co-information. For each of them, we consider its values over all MSA columns (or pairs or triplets of columns), and all amino acids if appropriate, for both natural and synthetic MSAs. To obtain the vertical and horizontal coordinates (respectively) of the markers in each panel, we compare these values for each natural MSA with the values from the corresponding synthetic MSAs generated by bmDCA or MSA Transformer (respectively). We use four different scores for this comparison, and devote each column of the figure to one of these scores – from the left: $|\rho - 1|$, where ρ denotes the Pearson correlation; $|\text{Slope} - 1|$, where "Slope" means the slope of best linear fit (see Fig. 4 for illustrations of these first two quantities in the case of two- and three-body connected correlations for family PF00153); the Jensen–Shannon divergence between the distributions of values; the Wasserstein distance between these distributions. For each statistical or information measure (row) and each score (column), and for each family in Table 1, we have one value of the score comparing the natural and bmDCA-generated MSAs and another one comparing the natural and MSA-Transformer–generated MSAs. We plot the former value versus the latter, yielding one marker per protein family in each plot. Thus, each plot compares the ability of bmDCA and MSA Transformer to reproduce the statistics of the natural data. Blue markers (above the diagonal) mean that the scores for MSA Transformer-generated MSAs are better, while green markers (below the diagonal) mean the opposite.

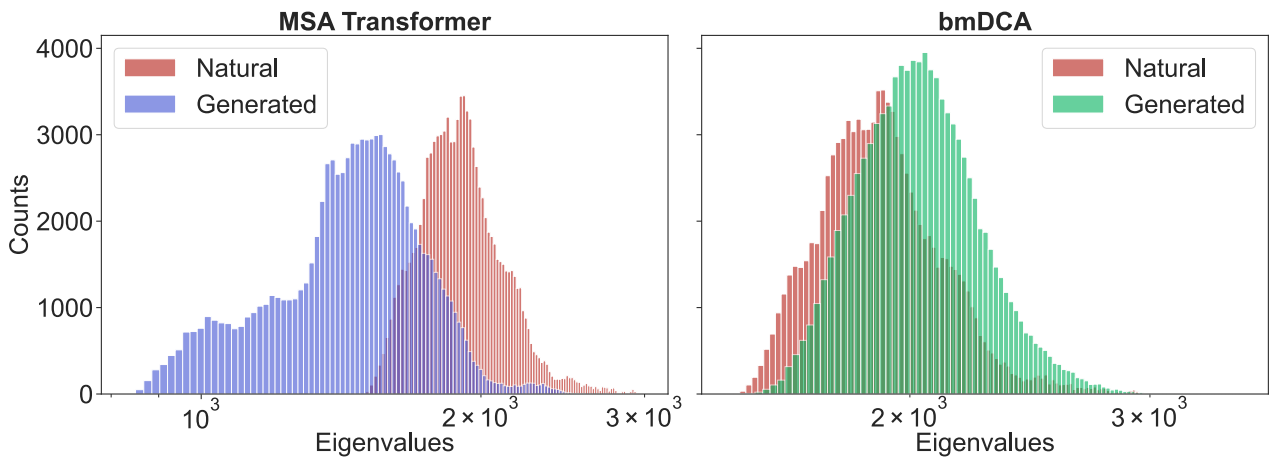

Figure S5: **Comparing phylogenies inferred from natural and generated MSAs for family PF00153.** We show the averaged spectra of modified graph Laplacian (MGL) matrices computed, for each tree inferred from an MSA, by using the leaves of multiple sub-trees made of 500 randomly sampled sequences from the MSA of interest. Specifically, in each case, we perform an average over 200 different sub-trees of the histograms of counts of the eigenvalues (see "Characterizing the distribution of sequences in MSAs"). We compare MSAs generated using either MSA Transformer (left) or bmDCA (right) to the natural MSA.