Protein language models trained on multiple sequence alignments learn phylogenetic relationships

Umberto Lupo[1,2,*], Damiano Sgarbossa[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
* Corresponding authors: umberto.lupo@epfl.ch, anne-florence.bitbol@epfl.ch

Abstract

Self-supervised neural language models with attention have recently been applied to biological sequence data, advancing structure, function and mutational effect prediction. Some protein language models, including MSA Transformer and AlphaFold's EvoFormer, take multiple sequence alignments (MSAs) of evolutionarily related proteins as inputs. Simple combinations of MSA Transformer's row attentions have led to state-of-the-art unsupervised structural contact prediction. We demonstrate that similarly simple, and universal, combinations of MSA Transformer's column attentions strongly correlate with Hamming distances between sequences in MSAs. Therefore, MSA-based language models encode detailed phylogenetic relationships. This could aid them to separate coevolutionary signals encoding functional and structural constraints from phylogenetic correlations arising from historical contingency. To test this hypothesis, we generate synthetic MSAs, either without or with phylogeny, from Potts models trained on natural MSAs. We demonstrate that unsupervised contact prediction is indeed substantially more resilient to phylogenetic noise when using MSA Transformer versus inferred Potts models.

Introduction

The explosion of available biological sequence data has led to multiple computational approaches aiming to infer three-dimensional structure, biological function, fitness, and evolutionary history of proteins from sequence data [1, 2]. Recently, self-supervised deep learning models based on natural language processing methods, especially attention [3] and transformers [4], have been trained on large ensembles of protein sequences by means of the masked language modeling objective of filling in masked amino acids in a sequence, given the surrounding ones [5–10]. These models, which capture long-range dependencies, learn rich representations of protein sequences, and can be employed for multiple tasks. In particular, they can predict structural contacts from single sequences in an unsupervised way [7], presumably by transferring knowledge from their large training set [11]. Neural network architectures based on attention are also employed in the "Evoformer" blocks in *AlphaFold* [12], as well as in *RoseTTAFold* [13] and *RGN2* [14], and they contributed to the recent breakthrough in the supervised prediction of protein structure.

Protein sequences can be classified in families of homologous proteins, that descend from an ancestral protein and share a similar structure and function. Analyzing multiple sequence alignments (MSAs) of homologous proteins thus provides substantial information about functional and structural constraints [1]. The statistics of MSA columns, representing amino-acid sites, allow to identify functional residues that are conserved during evolution, and correlations of amino acid usage between columns contain key information about functional sectors and structural contacts [15–18]. Indeed, through the course of evolution, contacting amino acids need to maintain their physico-chemical complementarity, which leads to correlated amino acid usages at these sites: this is known as coevolution. Potts models, also known as Direct Coupling Analysis (DCA), are pairwise maximum entropy models trained to match the empirical one- and two-body frequencies of amino acids observed in the columns

of an MSA of homologous proteins [2, 19–26]. They capture the coevolution of contacting amino-acids, and they provided state-of-the-art unsupervised predictions of structural contacts before the advent of protein language models. Note that coevolutionary signal also aids supervised contact prediction [27].

While most protein language neural networks take individual amino acid sequences as inputs, some others have been trained to perform inference from multiple sequence alignments (MSAs) of evolutionarily related sequences. This second class of networks includes *MSA Transformer* [28] and the "Evoformer" blocks in AlphaFold [12], both of which interleave per-sequence ("row") attention with per-site ("column") attention. Such an architecture is conceptually extremely attractive because it can incorporate coevolution in the framework of deep learning models using attention. In the case of MSA Transformer, simple combinations of the model's row attention heads have led to state-of-the-art unsupervised structural contact predictions, outperforming both language models trained on individual sequences and Potts models [28]. Beyond structure prediction, MSA Transformer is also able to predict mutational effects [29] and to capture fitness landscapes [30].

In addition to coevolutionary signal caused by structural and functional constraints, MSAs feature correlations that directly stem from the common ancestry of homologous proteins, i.e. from phylogeny. Does MSA Transformer learn to identify phylogenetic relationships between sequences, which are a key aspect of the MSA data structure? We demonstrate that simple, and universal, combinations of MSA Transformer's *column* attention heads, computed on a given MSA, strongly correlate with the Hamming distances between sequences in that MSA. This shows that MSA Transformer encodes detailed phylogenetic relationships. Does this help it to separate coevolutionary signals encoding functional and structural constraints from phylogenetic correlations arising from historical contingency? To address this question, we generate controlled synthetic MSAs from Potts models trained on natural MSAs, either without or with phylogeny. For this, we perform Metropolis Monte Carlo sampling under the Potts Hamiltonians, either at equilibrium or along phylogenetic trees inferred from the natural MSAs. Using the top Potts model couplings as proxies for structural contacts, we demonstrate that unsupervised contact prediction via MSA Transformer is substantially more resilient to phylogenetic noise than contact prediction using inferred Potts models.

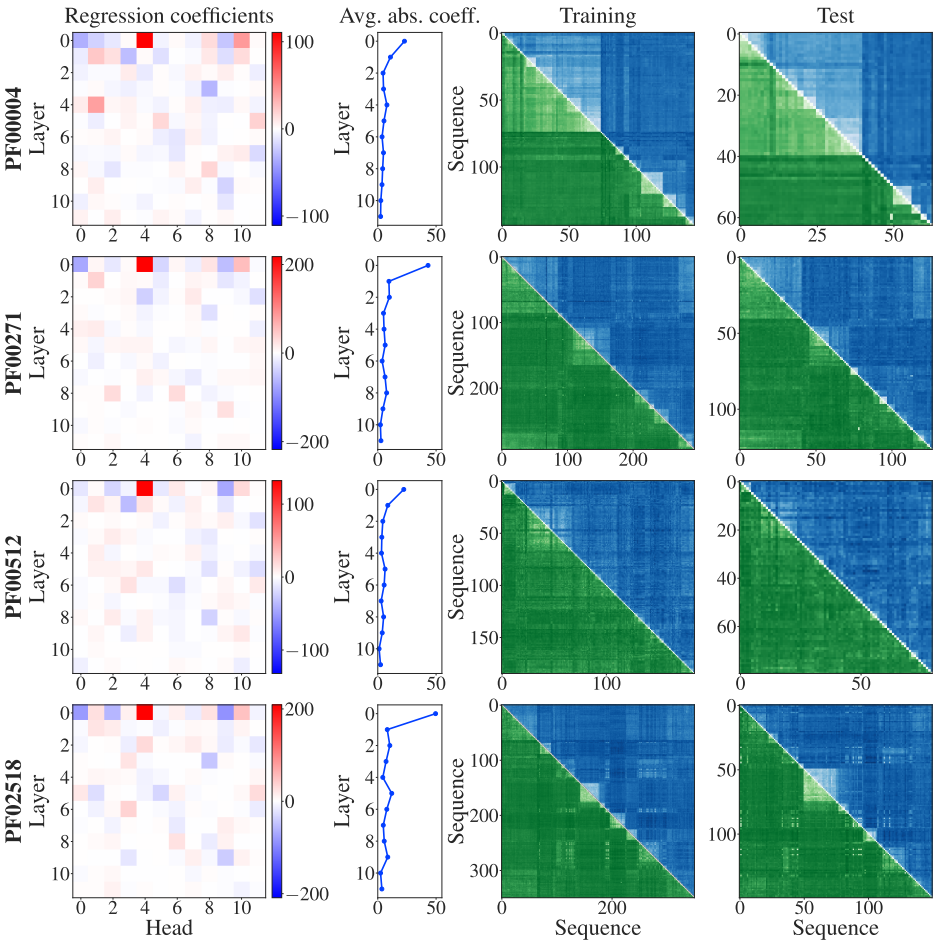
Results

Column attention heads capture Hamming distances in separate MSAs

We first considered separately each of 15 different Pfam seed MSAs (see "Datasets" and Table S1), corresponding to distinct protein families, and asked whether MSA Transformer has learned to encode phylogenetic relationships between sequences in its attention layers. To test this, we split each MSA randomly into a training and a test set, and train a logistic model [Eqs. (4) and (5)] based on the column-wise means of MSA Transformer's column attention heads on all pairwise distances in the training set – see "Supervised prediction of Hamming distances" for details. Fig. 1 shows the results of fitting these specialized logistic models.

For all alignments considered, large regression coefficients concentrate in early layers in the network, and single out some specific heads consistently across different MSAs – see Fig. 1(a), first and second columns, for results on four example MSAs. These logistic models reproduce the Hamming distances in the training set very well, and successfully predict those in the test set – see Fig. 1(a), third and fourth columns, for results on four example MSAs. Note that the block structures visible in the Hamming distance matrices, and well reproduced by our models, come from the phylogenetic ordering of sequences in our seed MSAs, see "Datasets". Quantitatively, in all the MSAs studied, the coefficients of determination (R^2) computed on the test sets are above 0.84 in all our MSAs – see Fig. 1(b).

A striking result from our analysis is that the regression coefficients appear to be similar across MSAs – see Fig. 1(a), first column. To quantify this, we computed the Pearson correlations between the regression coefficients learnt on the "larger" seed MSAs. Fig. 2 demonstrates that regression coefficients are indeed highly correlated across these MSAs.



(a)

Family	R^2
PF00004	0.97
PF00005	0.99
PF00041	0.98
PF00072	0.99
PF00076	0.98
PF00096	0.94
PF00153	0.95
PF00271	0.94
PF00397	0.84
PF00512	0.94
PF00595	0.98
PF01535	0.86
PF02518	0.92
PF07679	0.99
PF13354	0.99

(b)

Figure 1: **Fitting logistic models to predict Hamming distances separately in each MSA.** The column-wise means of MSA Transformer's column attention heads are used to predict normalised Hamming distances as "probabilities" in a logistic model. Each MSA is randomly split into a training set comprising 70% of its sequences and a test set composed of the remaining sequences. For each MSA, a logistic model is trained on all pairwise distances in the training set. (a) Regression coefficients are shown for each layer and attention head (first column), as well as their absolute values averaged over heads for each layer (second column). For four example MSAs, ground truth Hamming distances are shown in the upper triangle (blue) and predicted Hamming distances in the lower triangle and diagonal (green), for the training and test sets (third and fourth columns). Darker shades correspond to larger Hamming distances. (b) R^2 coefficients of determination for the predictions by each fitted model on its respective test set.

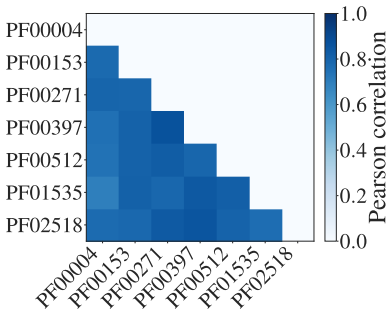


Figure 2: **Pearson correlations between regression coefficients in "larger" MSAs.** Sufficiently deep ($\geq$ 100 sequences) and long ($\geq$ 30 residues) MSAs are considered (mean/min/max Pearson correlations: 0.80/0.69/0.87).

MSA Transformer learns a universal representation of Hamming distances

Given the substantial similarities between our models trained separately on different MSAs, we next asked whether a common model across MSAs could capture Hamming distances within generic MSAs. To address this question, we trained a single logistic model, based on the column-wise means of MSA Transformer's column attention heads, on all pairwise distances within each of the first 12 of our seed MSAs. We assessed its ability to predict Hamming distances in the remaining 3 seed MSAs, which thus correspond to entirely different Pfam families from those in the training set. Fig. 3 shows the coefficients of this regression (first and second panels), as well as comparisons between predictions and ground truth values for the Hamming distances within the three test MSAs (last three panels). We observe that large regression coefficients again concentrate in the early layers of the model, but somewhat less than in individual models. Furthermore, the common model captures well the main features of the Hamming distance matrices in test MSAs.

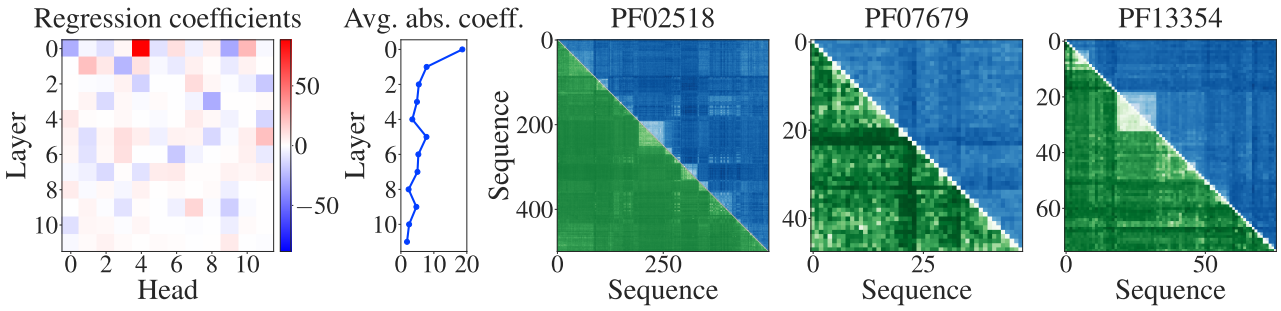


Figure 3: **Fitting a single logistic model to predict Hamming distances.** Our collection of 15 MSAs is split into a training set comprising 12 of them and a test set composed of the remaining 3. A logistic regression is trained on all pairwise distances within each MSA in the training set. Regression coefficients (first panel) and their absolute values averaged over heads for each layer (second panel) are shown as in Fig. 1. For the three test MSAs, ground truth Hamming distances are shown in the upper triangle (blue) and predicted Hamming distances in the lower triangle and diagonal (green), also as in Fig. 1 (last three panels).

In Table S2, we quantify the quality of fit for this model on all our MSAs. In all cases, we find very high Pearson correlation between the predicted distances and the ground truth Hamming distances. Furthermore, the median value of the R^2 coefficient of determination is 0.6, confirming the good quality of fit. In the three shortest and the two shallowest MSAs, the model performs below this median, while all MSAs for which R^2 is above median satisfy $M \geq 52$ and $L \geq 67$. We also compute, for each MSA, the slope of the linear fit when regressing the ground truth Hamming distances on the distances predicted by the model. MSA depth is highly correlated with the value of this slope (Pearson $r \approx 0.95$). This bias may be explained by the under-representation in the training set of Hamming distances and attention values from shallower MSAs, as their number is quadratic in MSA depth.

Ref. [28] showed that some column attention matrices, summed along one of their dimensions, correlate with phylogenetic sequence weights (see "Supervised prediction of Hamming distances"). This indicates that the model is, in part, attending to maximally diverse sequences. Our study demonstrates that MSA Transformer actually learns pairwise phylogenetic relationships between sequences, beyond these aggregate phylogenetic sequence weights. It also suggests an additional mechanism by which the model may be attending to these relationships, focusing on similarity instead of diversity. Indeed, while our regression coefficients with positive sign in Fig. 3 are associated with (average) attentions that are positively correlated with the Hamming distances, we also find several coefficients with large negative values. They indicate the existence of important *negative* correlations: in those heads, the model is actually attending to pairs of *similar* sequences. Besides, comparing our Figs. 1(a) and 3 with Fig. 5 in [28] shows that different attention heads are important in our study versus in the analysis of Ref. [28, Sec. 5.1]. Specifically, here we find that the fifth attention head in the first layer in the network is associated with the largest positive regression coefficient, while the sixth one

was most important there. Moreover, still focusing on the first layer of the network, the other most prominent heads here were not significant there.

For each layer and attention head in the network, MSA Transformer computes one matrix of column attention values per site present in the input MSA (plus one additional matrix corresponding to a "beginning-of-sentence" token). However, the common logistic model presented in this section was trained on the means of these matrices – see Eq. (3). To assess the contribution of each column, we employed this model to predict Hamming distances using column attentions coming from individual sites. We find that the most highly conserved sites (corresponding to columns with low entropy) lead to predictions whose errors have among the smallest standard deviations – see Table S3. Note that we focused on standard deviations to mitigate the biases of the common logistic model (see above). This indicates that highly conserved sites lead to more stable predictions.

MSA Transformer efficiently disentangles correlations from contacts and phylogeny

MSA Transformer is known to capture three-dimensional contacts through its (tied) row attention heads [28], and we have shown that it also captures Hamming distances, and thus phylogeny, through its column attention heads. How efficiently does MSA transformer disentangle correlations from contacts and phylogeny? We address this question in the concrete case of structure prediction. Because correlations from contacts and phylogeny are always both present in natural data, we constructed controlled synthetic data by sampling from Potts models, either independently at equilibrium, or along a phylogenetic tree inferred from the natural MSA using `FastTree 2` [31]. The Potts models we used were trained on each of 15 "full" natural MSAs (see "Datasets" and Table S1) using the generative method bmDCA [26, 32] – see "Synthetic MSA generation via Potts model sampling along inferred phylogenies". This setup allows us to compare data where all correlations come from couplings (pure Potts model) to data that comprises phylogenetic correlations on top of these couplings. For simplicity, let us call "contacts" the top scoring pairs of amino acid sites according to the bmDCA models used to generate our MSAs, and refer to the task of inferring these top scoring pairs as "contact prediction".

Contact maps inferred by plmDCA [24, 25] and by MSA Transformer for our synthetic datasets are shown in Fig. S3. For datasets generated with phylogeny, more false positives, scattered across the whole contact maps, appear in the inference by plmDCA than in that by MSA Transformer. This is shown quantitatively in Table 1, which reports the area under the receiver operating characteristic curve (ROC-AUC) for contact prediction for two different cutoffs on the number of contacts. We also quantify the degradation in performance caused by phylogeny by computing the relative drop Δ in ROC-AUC due to the injection of phylogeny in our generative process, for each Pfam family and for both plmDCA and MSA Transformer. On average, Δ is twice or three times (depending on the cutoff) higher for plmDCA than for MSA Transformer. We checked that these outcomes are robust to changes in the strategy used to compute plmDCA scores. In particular, the average Δ for plmDCA becomes even larger when we average scores coming from independent models fitted on the 10 subsampled MSAs used for MSA Transformer – thus using the exact same method as for predicting contacts with MSA Transformer (see "Generating sequences along an inferred phylogeny under a Potts model"). The conclusion is the same if 10 (or 6, for Pfam family PF13354) twice-deeper subsampled MSAs are employed.

These results demonstrate that contact inference by MSA Transformer is less deteriorated by phylogenetic correlations than contact inference by DCA. This analysis explains the remarkable result that structural contacts are predicted more accurately by MSA Transformer than by Potts models even when MSA Transformer's pre-training dataset minimizes diversity [28, Sec. 5.1].

An interesting feature that can be observed in Fig. S3 is that MSA Transformer tends to recover the experimental contact maps from our synthetic data generated by bmDCA. Specifically, some secondary features that were partially lost in the bmDCA inference and generation process (see the experimental contact maps in Fig. S2) become better defined again upon contact inference by MSA Transformer. This could be because MSA Transformer has learnt the structure of contact maps, including the spatial compactness and shapes of secondary structures.

	ROC-AUC for N contacts						ROC-AUC for $2L$ contacts					
	plmDCA			MSA Trans.			plmDCA			MSA Trans.		
Pfam ID	Eq.	Tree	Δ	Eq.	Tree	Δ	Eq.	Tree	Δ	Eq.	Tree	Δ
PF00004	0.87	0.58	0.33	0.70	0.67	0.04	0.93	0.61	0.34	0.80	0.71	0.11
PF00005	0.93	0.67	0.28	0.79	0.76	0.03	0.96	0.74	0.23	0.81	0.82	−0.01
PF00041	0.86	0.64	0.25	0.69	0.62	0.10	0.94	0.73	0.22	0.87	0.79	0.09
PF00072	0.94	0.73	0.23	0.86	0.77	0.10	0.99	0.85	0.14	0.94	0.87	0.08
PF00076	0.92	0.69	0.25	0.81	0.76	0.05	0.97	0.72	0.25	0.88	0.83	0.05
PF00096	0.88	0.54	0.39	0.68	0.54	0.21	0.92	0.54	0.41	0.78	0.54	0.30
PF00153	0.95	0.71	0.26	0.83	0.63	0.24	0.98	0.77	0.21	0.90	0.65	0.28
PF00271	0.91	0.62	0.32	0.78	0.72	0.07	0.95	0.67	0.29	0.85	0.77	0.10
PF00397	0.85	0.58	0.33	0.69	0.58	0.15	0.93	0.61	0.34	0.76	0.59	0.22
PF00512	0.94	0.74	0.21	0.84	0.77	0.08	0.97	0.78	0.20	0.88	0.81	0.08
PF00595	0.91	0.61	0.33	0.72	0.62	0.14	0.96	0.64	0.33	0.83	0.68	0.18
PF01535	0.85	0.66	0.23	0.66	0.63	0.05	0.88	0.72	0.18	0.73	0.72	0.01
PF02518	0.93	0.69	0.27	0.82	0.75	0.09	0.98	0.78	0.20	0.90	0.79	0.12
PF07679	0.85	0.63	0.26	0.68	0.64	0.05	0.95	0.77	0.19	0.85	0.80	0.05
PF13354	0.68	0.56	0.18	0.76	0.65	0.14	0.82	0.65	0.21	0.91	0.74	0.19
Average	0.88	0.64	0.27	0.75	0.68	0.10	0.94	0.71	0.25	0.85	0.74	0.12

Table 1: **Impact of phylogeny on contact prediction by plmDCA and MSA Transformer.** We consider synthetic MSAs generated by sampling Potts models either at equilibrium ("Eq.") or along inferred phylogenies ("Tree"). We report the ROC-AUCs for contact prediction, computed by comparing couplings inferred from our synthetic MSAs using plmDCA and MSA Transformer, with ground-truth proxy "contacts" consisting of either the N or the $2L$ pairs with top coupling scores according to the Potts models that generated the data (see "Synthetic MSA generation via Potts model sampling along inferred phylogenies"). Here, N denotes the number of pairs of residues that have an all-atom distance smaller than 8 Å in the experimental structure in Table S1, excluding pairs at positions i, j with $|i - j| \leq 4$ (in all cases, $N > 2L$). To assess the impact of phylogenetic noise, we compute $\Delta := (A_{\text{eq}} - A_{\text{tree}})/A_{\text{eq}}$, where A_{eq} is the ROC-AUC obtained from the equilibrium MSA and A_{tree} is the ROC-AUC obtained from the MSA with phylogeny.

Discussion

MSA Transformer is known to capture structural contacts through its (tied) row attention heads [28]. Here, we showed that it also captures Hamming distances, and thus phylogenetic information, through its column attention heads. This separation of the two signals in the representation of MSAs built by MSA Transformer comes directly from its architecture with inter-leaved row and attention heads. It makes sense, given that some correlations between columns (i.e. amino-acid sites) of an MSA are associated to contacts between sites, while similarities between rows (i.e. sequences) arise from relatedness between sequences [15]. Specifically, we found that simple combinations of column attention heads, tuned to individual MSAs, can predict pairwise Hamming distances between held-out sequences with very high accuracy. The larger coefficients in these combinations are found in early layers in the network. More generally, this study demonstrated that the regressions trained on different MSAs had major similarities. This motivated us to train a single model across a heterogeneous collection of MSAs, and this general model was still found to accurately predict pairwise distances in test MSAs from entirely distinct Pfam families. This result hints at a universal representation of phylogeny in MSA Transformer. Furthermore, our results suggest that the network has learned to quantify phylogenetic relatedness by attending not only to dissimilarity [28], but also to similarity relationships.

Next, to test the ability of MSA Transformer to disentangle phylogenetic correlations from functional and structural ones, we focused on unsupervised contact prediction tasks. Using controlled

synthetic data, we showed that unsupervised contact prediction is more robust to phylogeny when performed by MSA Transformer than by inferred Potts models.

Language models often capture important properties of the training data in their internal representations [33]. For instance, those trained on single protein sequences learn structure and binding sites [34], and those trained on chemical reactions learn how atoms rearrange [35]. Our finding that detailed phylogenetic relationships between sequences are learnt by MSA Transformer, in addition to structural contacts, and in an orthogonal way, demonstrates how precisely this model represents the MSA data structure. We note that, without language models, analyzing the correlations in MSAs can reveal evolutionary relatedness and sub-families [15], as well as collective modes of correlation, some of which are phylogenetic and some functional [18]. Furthermore, Potts models capture the clustered organization of protein families in sequence space [26], and the latent space of variational autoencoder models trained on sequences [36–38] qualitatively captures phylogeny [37]. Here, we demonstrated the stronger result that detailed pairwise phylogenetic relationships between sequences are quantitatively learnt by MSA Transformer.

Separating coevolutionary signals encoding functional and structural constraints from phylogenetic correlations arising from historical contingency constitutes a key problem in analyzing the sequence-to-function mapping in proteins [15, 18]. Phylogenetic correlations are known to obscure the identification of structural contacts by traditional coevolution methods, in particular by inferred Potts models [20, 21, 39–42], motivating various corrections [17, 21, 22, 24, 43–46]. From a theoretical point of view, disentangling these two types of signals is a fundamentally hard problem [47]. In this context, the fact that protein language models such as MSA Transformer learn both signals in orthogonal representations, and separate them better than Potts model, is remarkable.

Here, we have focused on Hamming distances as a simple measure of phylogenetic relatedness between sequences. It would be very interesting to extend our study to other, more detailed, measures of phylogeny. One may ask whether they are encoded in deeper layers in the network than those most involved in our study. Besides, we have mainly considered attentions averaged over columns, but exploring in more detail the role of individual columns would be valuable, especially given the impact we found for column entropies. More generally, the ability of protein language models to learn phylogeny raises the question of their possible usefulness to infer phylogenies and evolutionary histories.

Materials and methods

Datasets

The Pfam database [48] contains a large collection of related protein regions ("families"), typically associated to functional units called domains that can be found in multiple protein contexts. For each of its families, Pfam provides an expert-curated "seed" alignment that contains a representative set of sequences. In addition, Pfam provides deeper "full" alignments that are automatically built by searching against a large sequence database using a profile hidden Markov model (HMM) built from the seed alignments.

For this work, we considered 15 Pfam families, and for each we constructed (or retrieved, see below) one MSA from its seed alignment – henceforth referred to as the *seed MSA* – and one from its full alignment – henceforth referred to as the *full MSA*. The seed MSAs were created by first aligning Pfam seed alignments (Pfam version 35.0, Nov. 2021) to their HMMs using the `hmmalign` command from the HMMER suite (http://hmmer.org, version 3.3.2), and then removing columns containing only insertions or gaps. We retained the original Pfam "tree" ordering, with sequences ordered according to phylogeny inferred by FastTree [31]. In the case of family PF02518, out of the initial 658 sequences, we kept only the first 500 in order to limit the memory requirements of our computational experiments to less than 64 GB. Of the full MSAs, six (PF00153, PF00397, PF00512, PF01535, PF13354) were created from Pfam full alignments (Pfam version 34.0, Mar. 2021), removing columns containing only insertions or gaps, and finally removing sequences where 10% or more characters were gaps. The remaining nine full MSAs were retrieved from the online repository https://github.com/matteofigliuzzi/bmDCA

(publication date: Dec. 2017) and were previously considered in Ref. [26]. These alignments were constructed from full Pfam alignments from an earlier release of Pfam.

An 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 will refer to L as the MSA length, and to M as its depth. For all but one (PF13354) of our full MSAs, $M > 36000$. Despite their depth, however, our full MSAs include some highly similar sequences due to phylogenetic relatedness, a usual feature of large alignments of homologous proteins. We computed the effective depth [20] of each MSA $\mathcal{M}$ as

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

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 $\delta = 0.2$. While $M_{\text{eff}}^{(0.2)}/M$ can be as low as 0.06 for our full MSAs, this ratio is close to 1 for all seed MSAs: it is almost 0.83 for PF00004, and larger than 0.97 for all other families.

Finally, for each Pfam domain considered, we retrieved one experimental three-dimensional protein structure, corresponding to a sequence present in the full MSA, from the PDB (https://www.rcsb.org). All these structures were obtained by X-ray crystallography and have R-free values between 0.13 and 0.29. Information about our MSAs is summarized in Table S1.

All these families have been previously considered in the literature and shown to contain coevolutionary signal detectable by DCA methods [26], making our experiments on contact prediction readily comparable with previous results. While the precise choice of Pfam families is likely immaterial for our investigation of the column attention heads computed by MSA Transformer, our domains' short lengths are convenient in view of MSA Transformer's large memory footprint – which is $O(LM^2) + O(L^2)$.

MSA Transformer and column attention

We used the pre-trained MSA Transformer model introduced in Ref. [28], retrieved from the Python Package Index as `fair-esm 0.4.0`. We briefly recall that this model was trained, with a variant of the masked language modeling (MLM) objective [49], on 26 million MSAs constructed from UniRef50 clusters (March 2018 release), and contains 100 million trained parameters. The input to the model is an MSA with L columns and M rows. First, the model pre-pends a special "beginning-of-sentence" token to each row in the input MSA (this is common in language models inspired by the BERT architecture [49]). Then, each residue (or token) is embedded independently, via a learned mapping from the set of possible amino acid/gap symbols into $\mathbb{R}^d$ ($d = 768$). To these obtained embeddings, the model adds two kinds of learned [6] scalar *positional encodings* [50], designed to allow the model to distinguish between (a) different aligned positions (columns), and (b) between different sequence positions (rows).[1] The resulting collection of $M \times (L+1)$ d-dimensional vectors, viewed as an $M \times (L+1) \times d$ array, is then processed by a neural architecture consisting of 12 layers. Each layer is a variant of the *axial attention* [51] architecture, consisting of a multi-headed (12 heads) *tied row attention* block, followed by a multi-headed (12 heads) *column attention* block, and finally by a feed-forward network.[2] Tied row attention incorporates the expectation that 3D structure should be conserved amongst sequences in an MSA; we refer the reader to [28] for technical details. Column attention works as follows: let $X_j^{(l)}$ by the $M \times d$ matrix corresponding to "column" j in the $M \times (L+1) \times d$ array output by the row attention block in layer l with $l = 1, \ldots, 12$. At each layer l and each head $h = 1, \ldots, 12$, the model learns three $d \times d$ matrices[3] $W_{\text{Q}}^{(l,h)}$, $W_{\text{K}}^{(l,h)}$ and $W_{\text{V}}^{(l,h)}$, used to obtain three $M \times d$ matrices

$$Q_j^{(l,h)} = X_j^{(l)} W_{\text{Q}}^{(l,h)}, \quad K_j^{(l,h)} = X_j^{(l)} W_{\text{K}}^{(l,h)}, \quad V_j^{(l,h)} = X_j^{(l)} W_{\text{V}}^{(l,h)},$$

whose rows are referred to as "query", "key", and "value" vectors respectively. The column attention from MSA column $j \in \{0, \ldots, L\}$ (where $j = 0$ corresponds to the beginning-of-sentence token), at

[1] We remark that removing the latter kind was shown in [28] to have only limited impact.

[2] Both attention blocks, and the feed-forward network, are in fact preceded by layer normalization [52].

[3] These matrices, *mutatis mutandis*, could be of dimension $d \times d'$ with $d' \neq d$.

layer l, and from head h, is then the $M \times M$ matrix

$$A_j^{(l,h)} := \mathrm{softmax}_{\mathrm{row}} \left(\frac{Q_j^{(l,h)} K_j^{(l,h)\mathrm{T}}}{\sqrt{d}} \right), \tag{2}$$

where we denote by $\mathrm{softmax}_{\mathrm{row}}$ the application of $\mathrm{softmax}(\xi_1, \ldots \xi_d) = (e^{\xi_1}, \ldots, e^{\xi_d}) / \sum_{k=1}^{d} e^{\xi_k}$ to each row of a matrix independently, and by $(\cdot)^{\mathrm{T}}$ matrix transposition.[4]

Supervised prediction of Hamming distances

Row i of the column attention matrices $A_j^{(l,h)}$ in Eq. (2) consists of M positive weights summing to one – one weight per row index i' in the original MSA. According to the usual interpretation of the attention mechanism [3, 4], the role of these weights may be described as follows: When constructing a new internal representation (at layer l) for the row-i, column-j residue position, the network distributes its "focus", according to these weights, among the M available representation vectors associated with each MSA row-i', column-j residue position (including $i' = i$). Since row attention precedes column attention in the MSA Transformer architecture, we remark that, even at the first layer, the row-i', column-j representation vectors that are processed by that layer's column attention block can encode information about the entire row i' in the MSA.

In Ref. [28, Sec. 5.1], it was shown that, for some layers l and heads h, averaging the $M \times M$ column attention matrices $A_j^{(l,h)}$ in Eq. (2) from all MSA columns j, and then averaging the result along the first dimension, yields M-dimensional vectors whose entries correlate reasonably well with the "phylogenetic sequence weights" w_i defined in Eq. (1). Larger weights are, by definition, associated with less redundant sequences, and MSA diversity is known to be important for coevolution-based methods – particularly in structure prediction tasks. Thus, these correlations can be interpreted as suggesting that the model is, in part, explicitly attending to a maximally diverse set of sequences.

Beyond this, we hypothesize that MSA Transformer may have learned to quantify and exploit phylogenetic correlations in order to optimize its performance in the MLM training objective of "filling in" randomly masked residue positions. To investigate this, we set up regression tasks in which, to predict the Hamming distance y between the i-th and the i'-th sequence in an MSA $\mathcal{M}$ of length L, we used the entries $a_{i,i'}^{(l,h)}$ at position (i, i') (henceforth $a^{(l,h)}$ for brevity) from the 144 matrices

$$\boldsymbol{A}^{(l,h)} := \frac{1}{2(L+1)} \sum_{j=0}^{L} \left(A_j^{(l,h)} + A_j^{(l,h)\mathrm{T}} \right), \quad \text{with} \quad 1 \le l \le 12 \text{ and } 1 \le h \le 12. \tag{3}$$

These matrices are obtained by averaging, across all columns $j = 0, \ldots, L$, the symmetrised column attention maps $A_j^{(l,h)}$ computed by MSA Transformer, when taking $\mathcal{M}$ as input. We highlight that column $j = 0$, corresponding to the beginning-of-sentence token, is included in the average defining $\boldsymbol{A}^{(l,h)}$.

We fit fractional logit models via quasi-maximum likelihood estimation [53] using the `statsmodels` package [54]. Namely, we model the relationship between the Hamming distance y and the aforementioned symmetrised, and averaged, attention values $\boldsymbol{a} = (a^{(1,1)}, \ldots, a^{(12,12)})$, as

$$\mathbb{E}[y \,|\, \boldsymbol{a}] = G_{\beta_0, \boldsymbol{\beta}}(\boldsymbol{a}), \quad \text{with} \quad G_{\beta_0, \boldsymbol{\beta}}(\boldsymbol{a}) := \sigma\left(\beta_0 + \boldsymbol{a}\boldsymbol{\beta}^{\mathrm{T}}\right), \tag{4}$$

where $\mathbb{E}[\cdot \,|\, \cdot]$ denotes conditional expectation, $\sigma(x) = (1 + e^{-x})^{-1}$ is the standard logistic function, and the coefficients β_0 and $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_{144})$ are determined by maximising the sum of Bernoulli log-likelihoods

$$\ell(\beta_0, \boldsymbol{\beta} \,|\, \boldsymbol{a}, y) = y \log[G_{\beta_0, \boldsymbol{\beta}}(\boldsymbol{a})] + (1 - y) \log[1 - G_{\beta_0, \boldsymbol{\beta}}(\boldsymbol{a})], \tag{5}$$

[4] As in the standard Transformer architecture [4], these attention matrices are then used to compute $M \times d$ matrices $Z_j^{(l,h)} = A_j^{(l,h)} V_j^{(l,h)}$, one for each MSA column j and head h. Projecting the concatenation $Z_j^{(l,1)} \mid \cdots \mid Z_j^{(l,12)}$, a single $M \times d$ matrix $Z_j^{(l)}$ is finally obtained at layer l. The collection $(Z_j^{(l)})_{j=1,\ldots,L}$, thought of as an $M \times (L+1) \times d$ array, is then passed along to the feed-forward layer.

evaluated over a training set of observations of y and $\boldsymbol{a}$.[5] For simplicity, we refer to these fractional logit models simply as "logistic models".

Using data from our seed MSAs (cf. Table S1), we performed two types of regression tasks. In the first one, we randomly partitioned the set of row indices in each separate MSA $\mathcal{M}$ into two subsets $I_{\mathcal{M},\,\text{train}}$ and $I_{\mathcal{M},\,\text{test}}$, with $I_{\mathcal{M},\,\text{train}}$ containing 70% of the indices. We then trained and evaluated one model for each $\mathcal{M}$, using as training data the Hamming distances, and column attentions, coming from (unordered) pairs of indices in $I_{\mathcal{M},\,\text{train}}$, and as test data the Hamming distances, and column attentions, coming from pairs of indices in $I_{\mathcal{M},\,\text{test}}$. The second type of regression task was a single model fit over a training dataset consisting of all pairwise Hamming distances, and column attentions, from the first 12 of our 15 MSAs. We then evaluated this second model over a test set constructed in an analogous way from the remaining 3 MSAs.

Synthetic MSA generation via Potts model sampling along inferred phylogenies

To assess the performance of MSA Transformer at disentangling signals encoding functional and structural (i.e. fitness) constraints from phylogenetic correlations arising from historical contingency, we generated and studied controlled synthetic data. Indeed, disentangling fitness landscapes from phylogenetic history in natural data poses a fundamental challenge [47]. This makes it very difficult to assess the performance of a method at this task directly on natural data, because gold standards where the two signals are well-separated are lacking. We resolved this conundrum by generating synthetic MSAs according to well-defined dynamics such that the presence of phylogeny can be controlled.

First, we inferred unrooted phylogenetic trees from our full MSAs (see "Datasets"), using `FastTree 2.1` [31] with its default settings. Our use of FastTree is motivated by the depth of the full MSAs, which makes it computationally prohibitive to employ more precise inference methods. Deep MSAs are needed for the analysis described below, since it relies on accurately fitting Potts models.

Then, we fitted Potts models on each of these MSAs using bmDCA [26] (`https://github.com/ranganathanlab/bmDCA`) with its default hyperparameters. These include, in particular, regularization strengths for the Potts model fields and couplings, both set at $\lambda = 10^{-2}$. With the exception of family PF13354, we trained all models for 2000 iterations and stored the fields and couplings at the last iteration; in the case of PF13354, we terminated training after 1480 iterations. In all cases, we verified that, during training, the model's loss had converged. The choice of bmDCA is motivated by the fact that, as has been shown in [26, 32], model fitting on natural MSAs using Boltzmann machine learning yields Potts models with good generative power. This sets it apart from other DCA inference methods, especially pseudo-likelihood DCA (plmDCA) [24, 25], which is the DCA standard for contact prediction, but cannot faithfully reproduce empirical one- and two-body marginals, making it a poor choice of a generative model [26].

Using the phylogenetic trees and Potts models inferred from each full MSA, we generated synthetic MSAs without or with phylogeny, as we now explain. In the remainder of this subsection, let $\mathcal{M}$ denote an arbitrary MSA from our set of full MSAs, L its length, and M its depth.

Consider a sequence of L amino acid sites. We denote by $\alpha_i \in \{1, \ldots, q\}$ 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 $\vec{\alpha} = (\alpha_1, \ldots, \alpha_L)$ reads [2, 20]:

$$H(\vec{\alpha}) = -\sum_{i=1}^{L} h_i(\alpha_i) - \sum_{j=1}^{L} \sum_{i=1}^{j-1} e_{ij}(\alpha_i, \alpha_j)\,. \tag{6}$$

The parameters $h_i(\alpha_i)$ and $e_{ij}(\alpha_i, \alpha_j)$ are inferred from $\mathcal{M}$ by bmDCA [26, 32]. The Potts model probability distribution is then given by the Boltzmann distribution associated to the Hamiltonian H in Eq. (6):

$$P(\vec{\alpha}) = \frac{e^{-H(\vec{\alpha})}}{Z}\,, \tag{7}$$

[5] This setup is similar to logistic regression, but allows for the dependent variable to take real values between 0 and 1. It can be equivalently described as a generalized linear model with binomial family and logit link.

where Z is a constant ensuring normalization. In this context, an iteration step in the standard Metropolis–Hastings algorithm, for Markov Chain Monte Carlo (MCMC) sampling from P, consists of a proposed move ("mutation") in which a site i is chosen uniformly at random, and its state α_i may be changed into another state chosen uniformly at random. Each of these attempted mutations is accepted or rejected according to the Metropolis criterion, i.e. with probability

$$p = \max\left[1, \exp\left(-\Delta H\right)\right], \tag{8}$$

where ΔH is the difference in the value of H after and before the mutation.

Generating independent equilibrium sequences under a Potts model. To generate a synthetic MSAs without phylogeny from each $\mathcal{M}$, we performed equilibrium MCMC sampling from the Potts model with Hamiltonian H in Eq. (6), using the Metropolis–Hastings algorithm. Namely, we started from a set of M randomly and independently initialized sequences, and proposed a total number N of mutations on each sequence. 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 [26]. We thus used the value of N estimated by bmDCA at the end of training. This yielded a synthetic MSA of the same depth M as the original full MSA $\mathcal{M}$, composed of independent equilibrium sequences.

Generating sequences along an inferred phylogeny under a Potts model. We also generated synthetic data using MCMC sampling along our inferred phylogenetic trees [40]. We started from an equilibrium ancestor sequence sampled as explained above, and placed it at the root (note that, while FastTree roots its trees arbitrarily, root placement does not matter; see below). Then, this sequence was evolved by successive duplication (at each branching of the tree) and mutation events (along each branch). Mutations were again modeled using for acceptance the Metropolis criterion in Eq. (8) with the Hamiltonian in Eq. (6). As the length b of a branch gives the estimated number of substitutions that occurred per site along it [31], we generate data by making a number of accepted mutations on this branch equal to the integer closest to bL. Since we traversed the entire inferred tree in this manner, the resulting sequences at the leaves of the tree yield a synthetic MSA of the same depth as the original full MSA $\mathcal{M}$. Finally, we verified that the Hamming distances between sequences in these synthetic MSAs were reasonably correlated with those between corresponding sequences in the natural MSAs – see Fig. S1.

Because we start from an ancestral equilibrium sequence, and then employ the Metropolis criterion, all sequences in the phylogeny are equilibrium sequences. Thus, some of the correlations between the sequences at the leaves of the tree can be ascribed to the couplings in the Potts model, as in the case of independent equilibrium sequences described above. However, their relatedness adds extra correlations, arising from the historical contingency in their phylogeny. Note that separating these ingredients is extremely tricky in natural data [47], which motivates our study of synthetic data.

Our procedure for generating MSAs along a phylogeny is independent of the placement of the tree's root. Indeed, informally, a tree's root placement determines the direction of evolution; hence, root placement should not matter when evolution is a time-reversible process. That evolution via our mutations and duplications is a time-reversible process is a consequence of the fact that we begin with *equilibrium* sequences at the (arbitrarily chosen) root. More formally, for an irreducible Markov chain with transition matrix $\mathcal{P}$ and state space Ω, and for any $n \geq 1$, let $\mathrm{Markov}_n(\pi, \mathcal{P})$ denote the probability space of chains $(X_k)_{0\leq k\leq n}$ with initial distribution π on Ω. If π is the chain's stationary distribution and π satisfies detailed balance, then, for any number of steps $n \geq 1$, any chain $(X_k)_{0\leq k\leq n} \in \mathrm{Markov}_n(\pi, \mathcal{P})$ is reversible in the sense that $(X_{n-k})_{0\leq k\leq n} \in \mathrm{Markov}_n(\pi, \mathcal{P})$. In our case, since the Metropolis–Hastings algorithm constructs an irreducible Markov chain whose stationary distribution satisfies detailed balance, and since duplication events are also time-reversible constraints imposed at each branching node, all ensemble observables are independent of root placement as long as the root sequences are sampled from the stationary distribution.

Assessing performance degradation due to phylogeny in coupling inference. DCA methods and MSA Transformer both offer ways to perform unsupervised inference of structural contacts from MSAs of natural proteins. In the case of DCA, the established methodology [24–26] is to (1) learn fields and couplings [see Eq. (6)] by fitting the Potts model, (2) change the gauge to the zero-sum gauge, (3) compute the Frobenius norms, for all pairs of sites (i, j), of the coupling matrices $(e_{ij}(\alpha, \beta))_{\alpha,\beta}$, and finally (4) apply the *average product correction* (APC) [17], yielding a coupling score E_{ij}. Top scoring pairs of sites are then predicted as being contacts. In the case of MSA Transformer [28], a single logistic regression (shared across all possible input MSAs) was trained to regress contact maps from a sparse linear combination of the symmetrized and APC-corrected *row attention* heads (see "MSA Transformer and column attention").

We applied these inference techniques, normally used to predict structural contacts, on our *synthetic* MSAs generated without and with phylogeny (see above). As proxies for structural contacts, we used the pairs of sites with top coupling scores in the Potts models used to generate the MSAs. Indeed, when presented with our synthetic MSAs generated at equilibrium, DCA methods for fitting Potts models should recover the ranks of these coupling scores well. Hence, their performance in this task provide a meaningful baseline against which performance when a phylogeny was used to generate the data, as well as MSA Transformer's performance, can be measured.

As a DCA method to infer these coupling scores, we used plmDCA [24, 25] as implemented in the `PlmDCA` package (`https://github.com/pagnani/PlmDCA`), which is the state-of-the-art DCA method for contact inference. We fitted one plmDCA model per synthetic MSA, using default hyperparameters throughout; these include, in particular, regularization strengths set at $\lambda = 10^{-2}$ for both fields and couplings, and automatic estimation of the phylogenetic cutoff δ in Eq. (1). We verified that these settings led to good inference of structural contacts on the original full MSAs by comparing them to the PDB structures in Table S1 – see Fig. S2. For each synthetic MSA, we computed coupling scores E_{ij} for all pairs of sites.

While Potts models need to be fitted on deep MSAs to achieve good contact prediction, MSA Transformer's memory requirements are considerable even at inference time, and the average depth of the MSAs used to train MSA Transformer was 1192 [28]. Concordantly, we could not run MSA Transformer on any of the synthetic MSAs in their entirety. Instead, we subsampled each synthetic MSA 10 times, by selecting each time a number M_{sub} of row indices uniformly at random, without replacement. We used $M_{\text{sub}} \approx 380$ for family PF13354 due to its greater length, and $M_{\text{sub}} \approx 500$ for all other families. Then, we computed for each subsample a matrix of coupling scores using MSA Transformer's row attention heads and the logits from the aforementioned logistic regression. Finally, we averaged the resulting 10 matrices to obtain a single matrix of coupling scores.

Acknowledgments

The authors thank Mohammed AlQuraishi for inspiring discussions. 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] D. de Juan, F. Pazos, and A. Valencia. Emerging methods in protein co-evolution. *Nat Rev Genet*, 14(4):249–261, 2013.

[2] 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, March 2018.

[3] D. Bahdanau, K. Cho, and Y. Bengio. Neural Machine Translation by Jointly Learning to Align and Translate (ICLR 2015), 2014. arXiv: 1409.0473.

[4] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. Attention is all you need. *Advances in Neural Information Processing Systems*, 30:5998–6008, 2017.

[5] A. Elnaggar, M. Heinzinger, C. Dallago, G. Rehawi, Y. Wang, L. Jones, T. Gibbs, T. Feher, C. Angerer, M. Steinegger, D. Bhowmik, and B. Rost. ProtTrans: towards cracking the language of life's code through self-supervised learning, 2020. bioRxiv: 2020.07.12.199554.

[6] A. Rives, J. Meier, T. Sercu, S. Goyal, Z. Lin, J. Liu, D. Guo, M. Ott, C. L. Zitnick, J. Ma, and R. Fergus. Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *Proc. Natl. Acad. Sci. U.S.A.*, 118(15), 2021.

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

[8] K. Choromanski, V. Likhosherstov, D. Dohan, X. Song, A. Gane, T. Sarlos, P. Hawkins, J. Davis, A. Mohiuddin, L. Kaiser, D. Belanger, L. Colwell, and A. Weller. Rethinking attention with performers. In *International Conference on Learning Representations*, 2021.

[9] A. Madani, B. McCann, N. Naik, N. S. Keskar, N. Anand, R. R. Eguchi, P.-S. Huang, and R. Socher. ProGen: language modeling for protein generation, 2020. arXiv: 2004.03497.

[10] A. Madani, B. Krause, E. R. Greene, S. Subramanian, B. P. Mohr, J. M. Holton, J. L. Olmos, C. Xiong, Z. Z. Sun, R. Socher, J. S. Fraser, and N. Naik. Deep neural language modeling enables functional protein generation across families, 2021. bioRxiv: 2021.07.18.452833.

[11] N. Bhattacharya, N. Thomas, R. Rao, J. Dauparas, P. K. Koo, D. Baker, Y. S. Song, and S. Ovchinnikov. Interpreting Potts and Transformer Protein Models Through the Lens of Simplified Attention. *Pac Symp Biocomput*, 27:34–45, 2022.

[12] J. Jumper, R. Evans, A. Pritzel, T. Green, M. Figurnov, O. Ronneberger, K. Tunyasuvunakool, R. Bates, A. Žídek, A. Potapenko, A. Bridgland, C. Meyer, S. A. A. Kohl, A. J. Ballard, A. Cowie, B. Romera-Paredes, S. Nikolov, R. Jain, J. Adler, T. Back, S. Petersen, D. Reiman, E. Clancy, M. Zielinski, M. Steinegger, M. Pacholska, T. Berghammer, S. Bodenstein, D. Silver, O. Vinyals, A. W. Senior, K. Kavukcuoglu, P. Kohli, and D. Hassabis. Highly accurate protein structure prediction with AlphaFold. *Nature*, 596(7873):583–589, 2021.

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

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

[15] G. Casari, C. Sander, and A. Valencia. A method to predict functional residues in proteins. *Nat. Struct. Biol.*, 2(2):171–178, 1995.

[16] M. Socolich, S. W. Lockless, W. P. Russ, H. Lee, K. H. Gardner, and R. Ranganathan. Evolutionary information for specifying a protein fold. *Nature*, 437(7058):512–518, 2005.

[17] 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, 2008.

[18] N. Halabi, O. Rivoire, S. Leibler, and R. Ranganathan. Protein sectors: evolutionary units of three-dimensional structure. *Cell*, 138(4):774–786, 2009.

[19] 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.

[20] 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, 2009.

[21] 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.

[22] F. Morcos, A. Pagnani, B. Lunt, A. Bertolino, D. S. Marks, C. Sander, R. Zecchina, J. N. Onuchic, T. Hwa, and M. Weigt. Direct-coupling analysis of residue coevolution captures native contacts across many protein families. *Proc. Natl. Acad. Sci. U.S.A.*, 108(49):E1293–1301, 2011.

[23] 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, 2012.

[24] M. Ekeberg, C. Lovkvist, Y. Lan, M. Weigt, and E. Aurell. Improved contact prediction in proteins: using pseudolikelihoods to infer Potts models. *Phys. Rev. E*, 87(1):012707, 2013.

[25] M. Ekeberg, T. Hartonen, and E. Aurell. Fast pseudolikelihood maximization for direct-coupling analysis of protein structure from many homologous amino-acid sequences. *J. Comput. Phys.*, 276:341–356, 2014.

[26] M. Figliuzzi, P. Barrat-Charlaix, and M. Weigt. How Pairwise Coevolutionary Models Capture the Collective Residue Variability in Proteins? *Mol Biol Evol*, 35(4):1018–1027, April 2018.

[27] L. A. Abriata, G. E. Tamò, B. Monastyrskyy, A. Kryshtafovych, and M. Dal Peraro. Assessment of hard target modeling in CASP12 reveals an emerging role of alignment-based contact prediction methods. *Proteins*, 86 Suppl 1:97–112, March 2018.

[28] R. M. Rao, J. Liu, R. Verkuil, J. Meier, J. Canny, P. Abbeel, T. Sercu, and A. Rives. MSA transformer. Proceedings of Machine Learning Research, 139:8844–8856, 2021. M. Meila and T. Zhang, editors.

[29] J. Meier, R. Rao, R. Verkuil, J. Liu, T. Sercu, and A. Rives. Language models enable zero-shot prediction of the effects of mutations on protein function. In A. Beygelzimer, Y. Dauphin, P. Liang, and J. W. Vaughan, editors, *Advances in Neural Information Processing Systems*, 2021.

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

[31] M. N. Price, P. S. Dehal, and A. P. Arkin. Fasttree 2 – approximately maximum-likelihood trees for large alignments. *PLOS ONE*, 5(3):1–10, March 2010.

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

[33] A. Rogers, O. Kovaleva, and A. Rumshisky. A primer in BERTology: what we know about how BERT works. *Transactions of the Association for Computational Linguistics*, 8:842–866, 2020.

[34] J. Vig, A. Madani, L. R. Varshney, C. Xiong, richard socher, and N. Rajani. BERTology meets biology: interpreting attention in protein language models. In *International Conference on Learning Representations*, 2021.

[35] P. Schwaller, B. Hoover, J. L. Reymond, H. Strobelt, and T. Laino. Extraction of organic chemistry grammar from unsupervised learning of chemical reactions. *Sci Adv*, 7(15), 2021.

[36] A. J. Riesselman, J. B. Ingraham, and D. S. Marks. Deep generative models of genetic variation capture the effects of mutations. *Nat Methods*, 15(10):816–822, October 2018.

[37] X. Ding, Z. Zou, and C. L. Brooks Iii. Deciphering protein evolution and fitness landscapes with latent space models. *Nat Commun*, 10(1):5644, December 2019.

[38] F. McGee, S. Hauri, Q. Novinger, S. Vucetic, R. M. Levy, V. Carnevale, and A. Haldane. The generative capacity of probabilistic protein sequence models. *Nat Commun*, 12(1):6302, November 2021.

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

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

[41] E. Rodriguez Horta, P. Barrat-Charlaix, and M. Weigt. Toward inferring potts models for phylogenetically correlated sequence data. *Entropy*, 21(11), 2019.

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

[43] O. Lichtarge, H. R. Bourne, and F. E. Cohen. An evolutionary trace method defines binding surfaces common to protein families. *J Mol Biol*, 257(2):342–358, 1996.

[44] A. J. Hockenberry and C. O. Wilke. Phylogenetic weighting does little to improve the accuracy of evolutionary coupling analyses. *Entropy*, 21(10), 2019.

[45] D. Malinverni and A. Barducci. Coevolutionary Analysis of Protein Subfamilies by Sequence Reweighting. *Entropy*, 21(11):1127, 2020.

[46] 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.

[47] E. N. Weinstein, A. N. Amin, J. Frazer, and D. S. Marks. Non-identifiability and the blessings of misspecification in models of molecular fitness and phylogeny. *bioRxiv*, 2022.

[48] J. Mistry, S. Chuguransky, L. Williams, M. Qureshi, G. Salazar, E. L. L. Sonnhammer, S. C. E. Tosatto, L. Paladin, S. Raj, L. J. Richardson, R. D. Finn, and A. Bateman. Pfam: The protein families database in 2021. *Nucleic Acids Research*, 49(D1):D412–D419, October 2020. eprint: https://academic.oup.com/nar/article-pdf/49/D1/D412/35363969/gkaa913.pdf.

[49] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics, June 2019.

[50] J. Gehring, M. Auli, D. Grangier, D. Yarats, and Y. N. Dauphin. Convolutional sequence to sequence learning. In D. Precup and Y. W. Teh, editors, *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 1243–1252. PMLR, 2017.

[51] J. Ho, N. Kalchbrenner, D. Weissenborn, and T. Salimans. Axial attention in multidimensional transformers, 2019. arXiv: 1912.12180.

[52] J. L. Ba, J. R. Kiros, and G. E. Hinton. Layer normalization, 2016. arXiv: 1607.06450.

[53] L. E. Papke and J. M. Wooldridge. Econometric methods for fractional response variables with an application to 401(k) plan participation rates. *J. Appl. Econ.*, 11(6):619–632, 1996.

[54] S. Seabold and J. Perktold. Statsmodels: econometric and statistical modeling with python. In *9th Python in Science Conference*, 2010.

Supplementary material

		Seed MSA		Full MSA			PDB structure	
Pfam ID	Family name	L	M	L	M	$M_{\text{eff}}^{(0.2)}$	ID	Resol.
PF00004	AAA	132	207	132	39277	9050	4D81	2.40 Å
PF00005	ABC_tran	137	55	137	68891	43882	1L7V	3.20 Å
PF00041	fn3	85	98	85	42721	17783	3UP1	2.15 Å
PF00072	Response_reg	112	52	112	73063	40180	3ILH	2.59 Å
PF00076	RRM_1	68	70	69	51964	20276	3NNH	2.75 Å
PF00096	zf-C2H2	23	159	23	38996	12581	4R2A	1.59 Å
PF00153	Mito_carr	97	160	94	93776	17860	1OCK	2.20 Å
PF00271	Helicase_C	111	421	111	66809	25018	3EX7	2.30 Å
PF00397	WW	31	448	31	39045	3361	4REX	1.60 Å
PF00512	HisKA	67	265	66	154998	67303	3DGE	2.80 Å
PF00595	PDZ	82	44	82	71303	4053	1BE9	1.82 Å
PF01535	PPR	31	458	31	109064	37514	4M57	2.86 Å
PF02518	HATPase_c	112	500	111	80714	59190	3G7E	2.20 Å
PF07679	I-set	90	48	90	36141	14611	1FHG	2.00 Å
PF13354	Beta-lactamase2	215	76	198	4642	3535	6QW8	1.10 Å

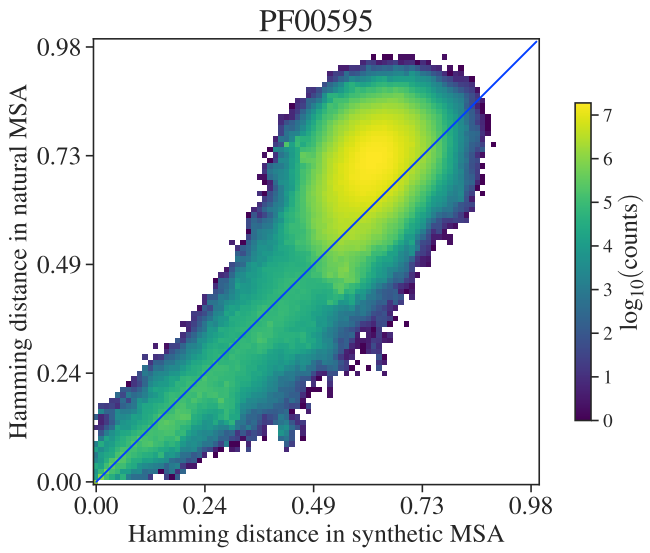
Table S1: **Pfam families and MSAs considered in this work.** For each family, we considered a shallow MSA constructed from the corresponding Pfam seed alignment ("Seed MSA") and a deep MSA constructed from the Pfam full alignment ("Full MSA") – see "Datasets". For both kinds of MSA, we report the length L and depth M. For the full MSAs, we report the effective depth $M_{\text{eff}}^{(0.2)}$ as defined in Eq. (1). The PDB structures used, and their resolutions, are also reported. Note that occasional length mismatches between seed and full MSAs reflect our use of data from a more recent Pfam release in the case of the seed MSAs.

Family	R^2	Pearson	Slope
PF00004	0.84	0.95	0.95
PF00005	0.72	0.92	0.75
PF00041	0.56	0.90	0.75
PF00072	0.66	0.90	0.71
PF00076	0.59	0.88	0.68
PF00096	0.57	0.88	0.73
PF00153	0.81	0.93	0.80
PF00271	0.77	0.93	1.11
PF00397	0.23	0.84	1.13
PF00512	0.77	0.93	0.94
PF00595	0.50	0.89	0.63
PF01535	0.54	0.86	1.18
PF02518	**0.60**	**0.90**	**1.20**
PF07679	**0.28**	**0.85**	**0.57**
PF13354	**0.67**	**0.92**	**0.70**

Table S2: **Quality of fit for our logistic model trained on Hamming distances and column attentions from several MSAs.** For the logistic model described in "MSA Transformer learns a universal representation of Hamming distances", and for the MSAs in the training set (plain font) and test set (boldface font), we report (1) the R^2 coefficient of determination for the model's predictions, (2) the Pearson correlation coefficient between predictions and ground truth Hamming distances, and (3) the slope of the line of best fit when regressing the ground truth Hamming distances on the model's predictions.

Family	Pearson	Z-score
PF00004	0.36	-2.29
PF00005	0.46	-2.78
PF00041	0.38	-2.21
PF00072	0.41	-1.60
PF00076	0.53	-1.64
PF00096	0.88	-1.71
PF00153	0.66	-3.06
PF00271	0.83	-2.24
PF00397	0.79	-1.21
PF00512	0.66	-1.73
PF00595	0.13	0.28
PF01535	0.89	-1.68
PF02518	0.81	-2.48
PF07679	0.66	-2.20
PF13354	0.39	-2.96

Table S3: **Relation between entropy and test error when using parameters from our common logistic model on individual column attention matrices.** We applied the common logistic model described in "MSA Transformer learns a universal representation of Hamming distances" – which was trained on the column-wise means of MSA Transformer's column attention heads [see Eq. (3)] – to individual column attention heads. We then computed the resulting errors in the prediction of Hamming distances. For each of our seed MSAs, we report the Pearson correlation between the entropy of each column and the standard deviation of the distribution of these errors. We also computed the mean of this standard deviation when restricting to the 5 columns with lowest entropy. Z-scores for this mean are reported, showing that these 5 columns have significantly lower standard deviations than the rest (except for PF00595).



(a)

Family	Pearson correlation	
	All	$d_{\text{nat}} \leq 0.5$
PF00004	0.72	0.75
PF00005	0.31	0.62
PF00041	0.31	0.90
PF00072	0.33	0.85
PF00076	0.25	0.84
PF00096	0.32	0.23
PF00153	0.39	0.91
PF00271	0.37	0.70
PF00397	0.60	0.73
PF00512	0.30	0.65
PF00595	0.67	0.96
PF01535	0.15	0.70
PF02518	0.39	0.67
PF07679	0.19	0.90
PF13354	0.74	0.79

(b)

Figure S1: **Comparing Hamming distances from synthetic MSAs generated along inferred phylogenies with Hamming distances from natural MSAs.** The Hamming distances between sequences in our synthetic MSAs generated along phylogenies (see "Generating sequences along an inferred phylogeny under a Potts model") are reasonably correlated with those between corresponding sequences in the natural MSAs used to infer the phylogenies. (a) Density plot comparing the Hamming distances between sequences in the synthetic MSA corresponding to family PF00595, with those between corresponding sequences in the natural MSA for this family. The pixel with coordinates $(d_{\text{synth}}, d_{\text{nat}})$ is colored according to the base-10 logarithm of the number of pairs of indices (i, j) such that the Hamming distance between row i and row j in the synthetic MSA is d_{synth}, and the distance between the same rows in the natural MSA is d_{nat}. Recall that the sequence at row i of the synthetic MSA has been generated on the leaf of the inferred phylogenetic tree that corresponds to the natural sequence at row i of the natural MSA. As is visible here, our synthetic MSAs tend to have slightly smaller Hamming distances, on average, than their natural counterparts. (b) For each Pfam family, we computed two Pearson correlation coefficients: first, the correlation between all Hamming distances among synthetic sequences and all Hamming distances among natural sequences; second, the same correlation but restricting to pairs of indices (i, j) such that the distance between sequence i and sequence j in the natural MSA is no larger than 0.5.

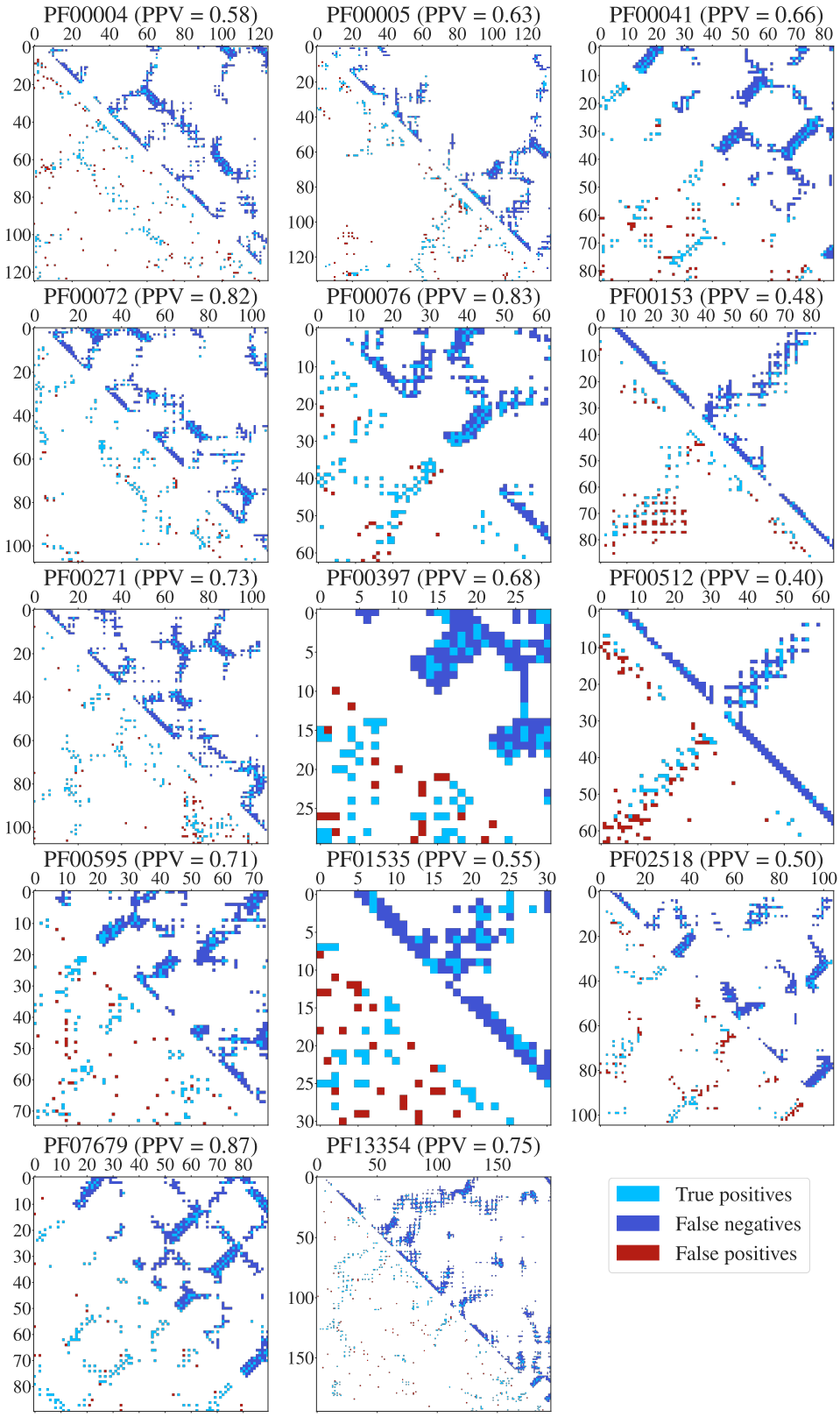


Figure S2: **Predicted contact maps using plmDCA versus experimental contact maps for 14 of our full natural MSAs.** Experimental contact maps are displayed in the upper-triangular portions of each panel, and are obtained from the PDB structures in Table S1 by using an all-atom Euclidean distance cutoff of 8 Å, excluding residue pairs at positions i, j with $|i - j| \leq 4$. Predictions are displayed in the lower-triangular portions, and are obtained considering the top $2L$ scores, where L is the length of the MSA. Light blue squares represent true positive predictions, dark blue squares false negative predictions, and red squares false positive predictions. For each predicted contact map, we report the positive predictive value (PPV) given these choices. Results for PF00096 are not displayed due to its very short length.

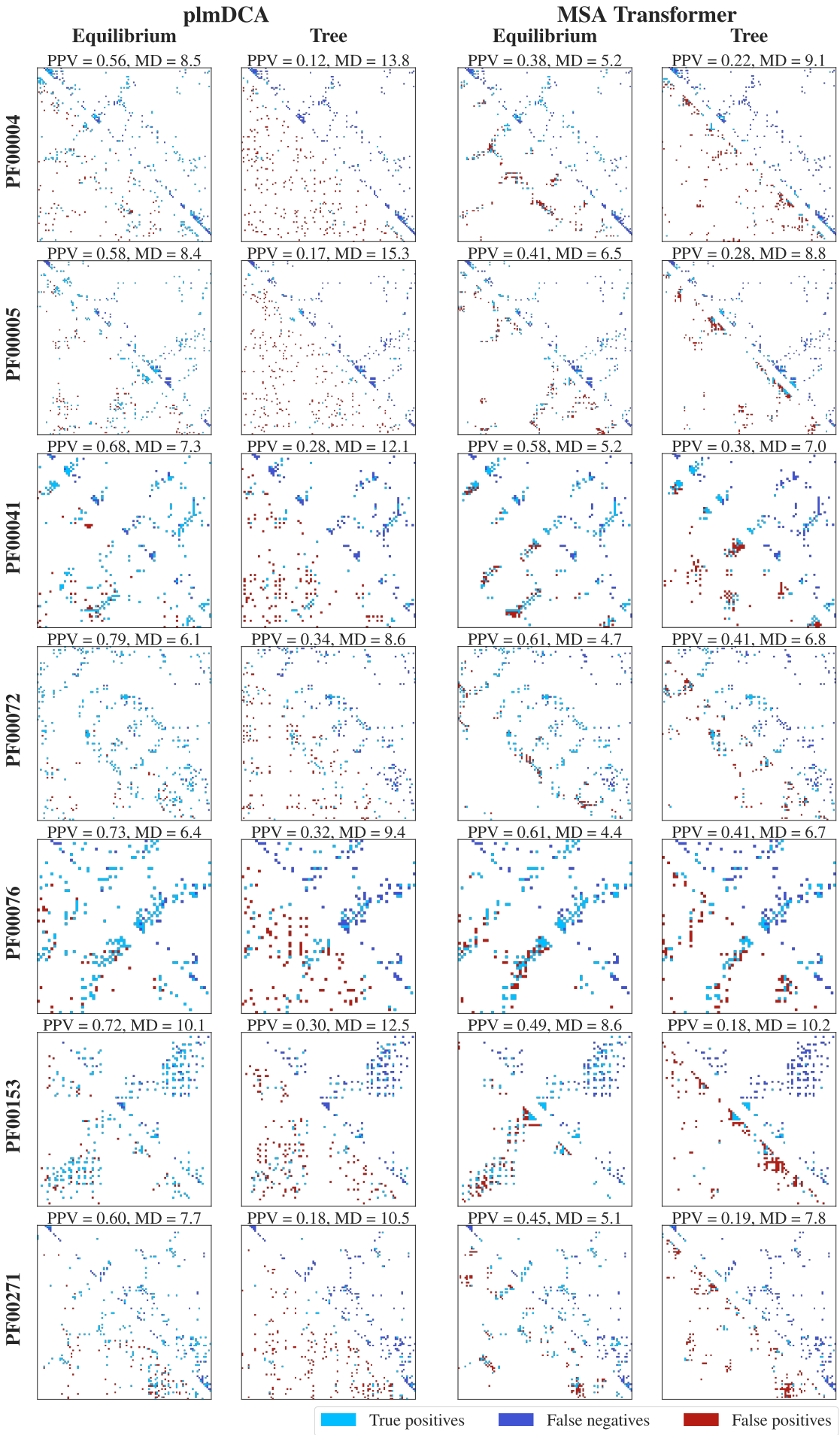


Figure S3: (Continued on the next page.)

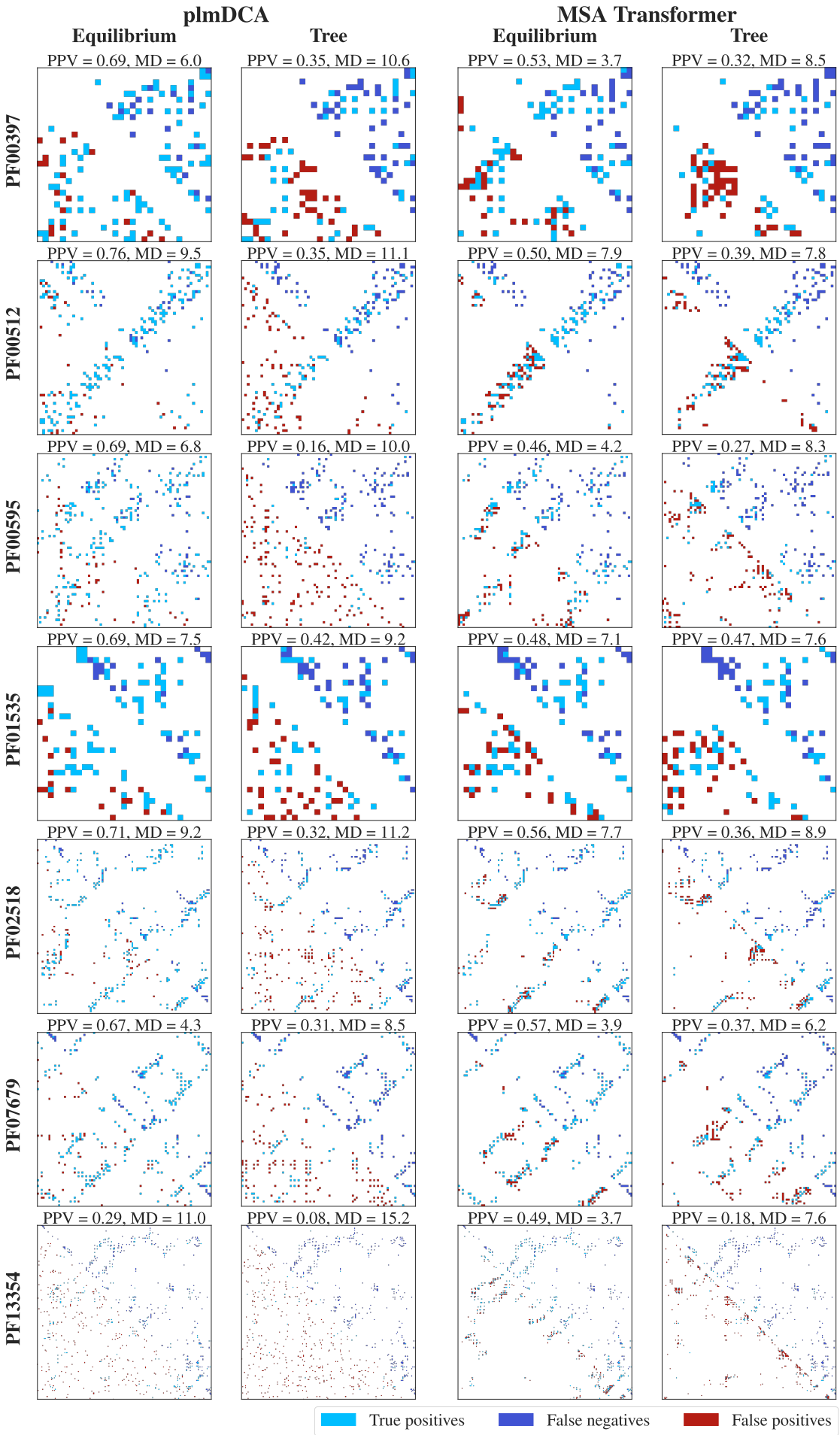


Figure S3: (Continued on the next page.)

Figure S3: **Predicted contact maps using plmDCA and MSA Transformer on synthetic MSAs, versus ground truth "contact maps" defined by top Potts model couplings, for 14 of our full MSAs.** "Contact maps" containing $2L$ contacts, and obtained from the ground-truth couplings in the Potts models used to generate our synthetic MSAs (see "Synthetic MSA generation via Potts model sampling along inferred phylogenies"), are displayed in the upper-triangular portions of each panel. In the lower-triangular portions, we display the $2L$ top-scoring pairs according to plmDCA or MSA Transformer, when performing contact inference on synthetic MSAs generated from those Potts models either without phylogeny ("Equilibrium") or with phylogeny ("Tree"). Light blue squares represent true positive predictions, dark blue squares false negative predictions, and red squares false positive predictions. For each predicted contact map, we report the positive predictive value (PPV) given these choices, as well as the median distance (MD), expressed in Å, between predicted pairs in the reference experimental 3D structure (see Table S1). ROC-AUC values corresponding to this analysis are reported in the columns titled "ROC-AUC for $2L$ contacts" of Table 1.