nature communications

Article

https://doi.org/10.1038/s41467-022-34032-y

Protein language models trained on multiple sequence alignments learn phylogenetic relationships

Received: 8 April 2022

Accepted: 7 October 2022

Published online: 22 October 2022

Check for updates

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

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. We further show that these models can separate coevolutionary signals encoding functional and structural constraints from phylogenetic correlations reflecting historical contingency. To assess this, we generate synthetic MSAs, either without or with phylogeny, from Potts models trained on natural MSAs. We find that unsupervised contact prediction is substantially more resilient to phylogenetic noise when using MSA Transformer versus inferred Potts models.

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

[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. e-mail: umberto.lupo@epfl.ch; anne-florence.bitbol@epfl.ch

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 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 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 row (i.e. per-sequence) attention with column (i.e. per-site) 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 prediction, outperforming both language models trained on individual sequences and Potts models[28]. Beyond structure prediction, MSA Transformer is also able to predict mutational effects[29,30] and to capture fitness landscapes[31]. 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?

Here, we show 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 demonstrates that MSA Transformer encodes detailed phylogenetic relationships. Is MSA Transformer able 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 "Methods – Datasets" and Supplementary Table 1), 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. (5) and (6)] based on the column-wise means of MSA Transformer's column attention heads on all pairwise Hamming distances in the training set—see Fig. 1 for a schematic, and "Methods – Supervised prediction of Hamming distances" for details. Figure 2 and Table 1 show 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. 2, 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. 2, 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 "Methods – 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 Table 1.

A striking result from our analysis is that the regression coefficients appear to be similar across MSAs—see Fig. 2, first column. To quantify this, we computed the Pearson correlations between the regression coefficients learnt on the larger seed MSAs. Figure 3 demonstrates that regression coefficients are indeed highly correlated across these MSAs.

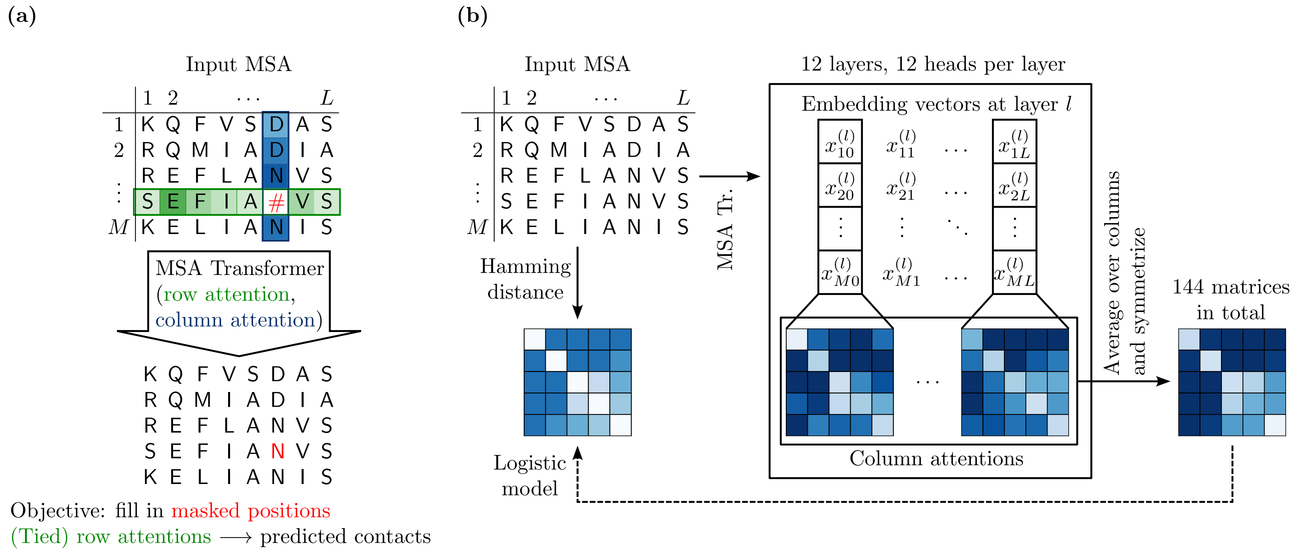

Fig. 1 | MSA Transformer: column attentions and Hamming distances. a MSA Transformer is trained using the masked language modeling objective of filling in randomly masked residue positions in MSAs. For each residue position in an input MSA, it assigns attention scores to all residue positions in the same row (sequence) and column (site) in the MSA. These computations are performed by 12 independent row/column attention heads in each of 12 successive layers of the network. **b** Our approach for Hamming distance matrix prediction from the column attentions computed by the trained MSA Transformer model, using a natural MSA as input. For each $i = 1, \ldots, M$, $j = 0, \ldots, L$ and $l = 1, \ldots, 12$, the embedding vector $x_{ij}^{(l)}$ is the i-th row of the matrix $X_j^{(l)}$ defined in "Methods – MSA Transformer and column attention", and the column attentions are computed according to Eqs. (2) and (3).

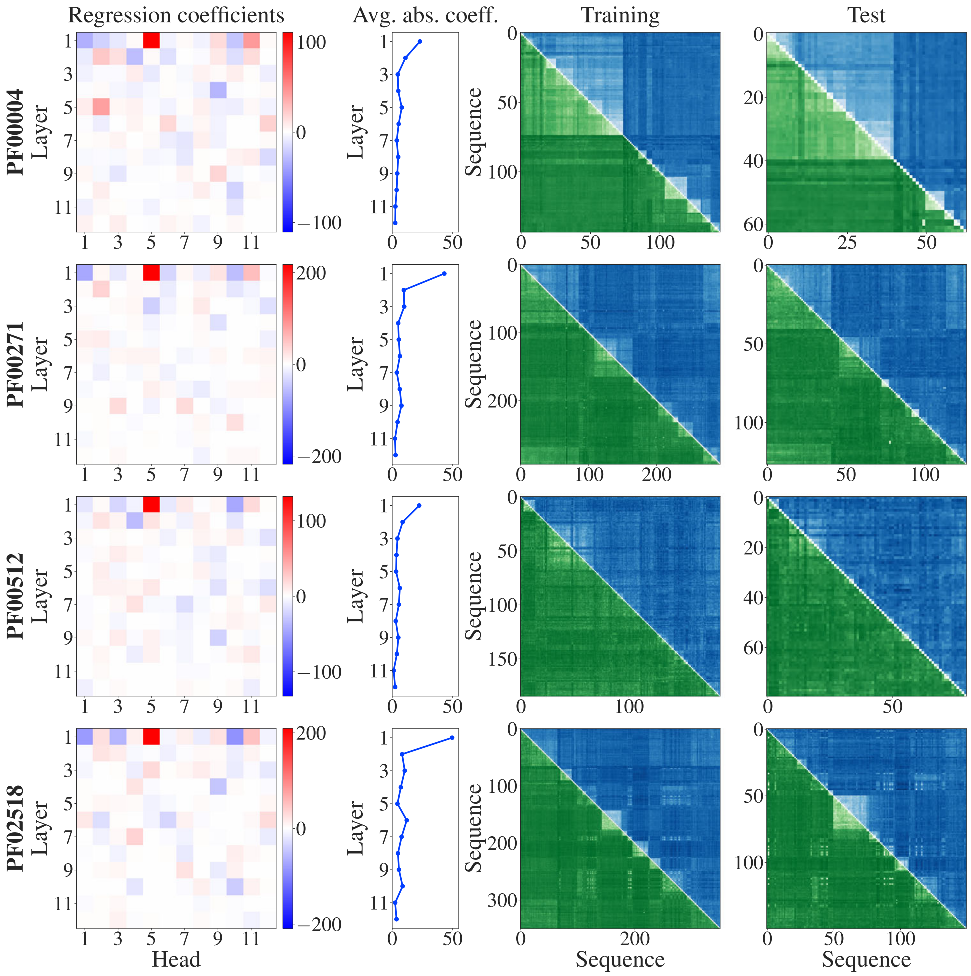

Fig. 2 | Fitting logistic models to predict Hamming distances separately in each MSA. The column-wise means of MSA Transformer's column attention heads are used to predict normalised Hamming distances as probabilities in a logistic model. Each MSA is randomly split into a training set comprising 70% 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. 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.

Table 1 | Quality of fit for logistic models trained to predict Hamming distances separately in each MSA

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

R^2 coefficients of determination are shown for the predictions by each fitted model on the associated test set, see Fig. 2.

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. Figure 4 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.

In Supplementary Table 2, 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 have depth $M \geq 52$ and length $L \geq 67$. 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.

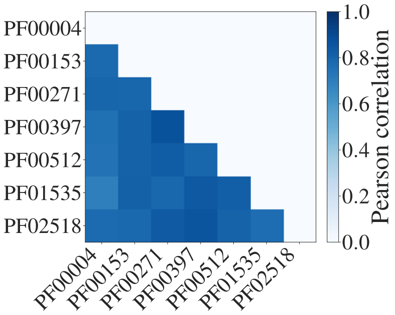

Fig. 3 | Pearson correlations between regression coefficients in larger MSAs. Sufficiently deep (≥ 100 sequences) and long (≥ 30 residues) MSAs are considered (mean/min/max Pearson correlations: 0.80/0.69/0.87).

Ref. 28 showed that some column attention matrices, summed along one of their dimensions, correlate with phylogenetic sequence weights (see "Methods – 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. 4 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. 2, 4 with Fig. 5 in ref. 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. MSA Transformer's ability to focus on similarity may also explain why its performance at predicting mutational effects can decrease significantly when using MSAs which include a duplicate of the query sequence (see ref. 29,

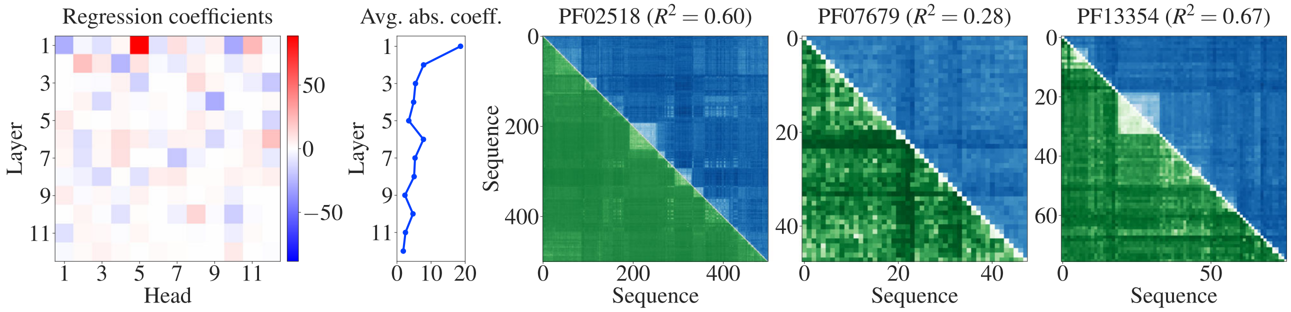

Fig. 4 | Fitting a single logistic model to predict Hamming distances. Our collection of 15 MSAs is split into a training set comprising 12 of them and a test set composed of the remaining 3. A logistic regression is trained on all pairwise distances within each MSA in the training set. Regression coefficients (first panel) and their absolute values averaged over heads for each layer (second panel) are shown as in Fig. 2. 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. 2 (last three panels). We further report the R^2 coefficients of determination for the regressions on these test MSAs—see also Supplementary Table 2.

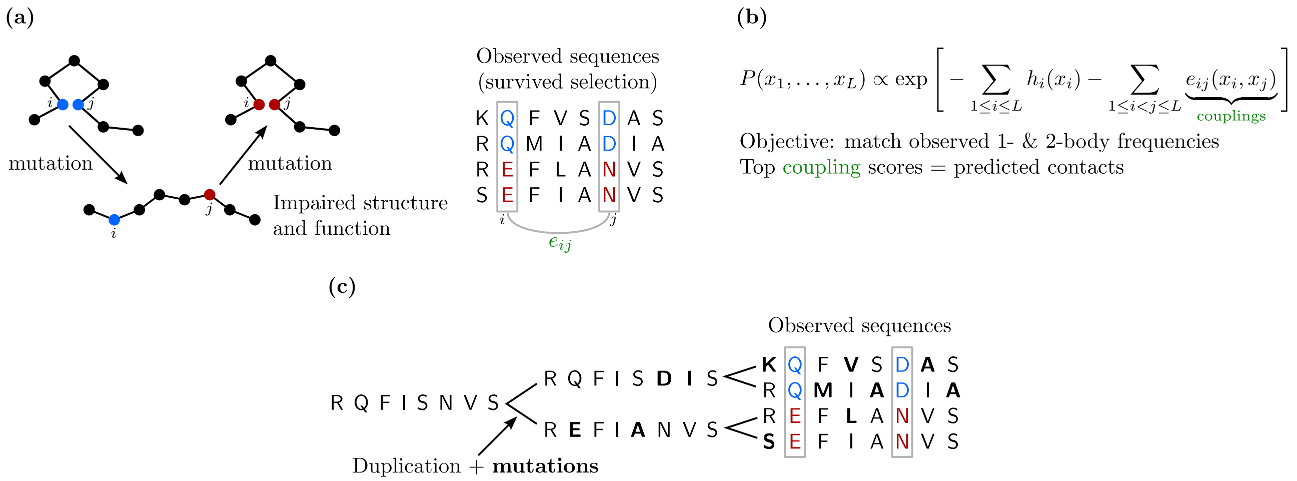

Fig. 5 | Correlations from coevolution and from phylogeny in MSAs. a Natural selection on structure and function leads to correlations between residue positions in MSAs (coevolution). **b** Potts models, also known as DCA, aim to capture these correlations in their pairwise couplings. **c** Historical contingency can lead to correlations even in the absence of structural or functional constraints.

Supplementary Fig. 9 and Table 10): in these cases, the model predicts masked tokens with very high confidence using information from the duplicate sequence.

How much does the ability of MSA Transformer to capture phylogenetic relationships arise from its training? To address this question, we trained a common logistic model as above to predict Hamming distances, but using column attention values computed from a randomly re-initialized version of the MSA Transformer network. We used the same protocol as in MSA Transformer's original pre-training to randomly initialize the entries of the network's row- and column-attention weight matrices $W_Q^{(l,h)}$, $W_K^{(l,h)}$ and $W_V^{(l,h)}$ (see "Methods – MSA Transformer and column attention"), as well as the entries of the matrix used to embed input tokens, the weights in the feed-forward layers, and the positional encodings. Specifically, we sampled these entries (with the exception of bias terms and of the embedding vector for the padding token, which were set to zero) from a Gaussian distribution with mean 0 and standard deviation 0.02. The results obtained in this case for our regression task are reported in Supplementary Table 3. They demonstrate that, although random initialization can yield better performance than random guessing (which may partly be explained by Gordon's Theorem[32]), the trained MSA Transformer gives vastly superior results. This confirms that the masked language modeling pre-training has driven it towards precisely encoding distances between sequences.

For each layer and attention head in the network, MSA Transformer computes one matrix of column attention values per site—see Eq. (4). This is in contrast with row attention, which is tied (see "Methods – MSA Transformer and column attention"). Our results are more surprising that they would be if the model's column attentions were also tied. Indeed, during pre-training, by tuning its row-attention weight matrices to achieve optimal tied attention, MSA Transformer discovers covariance between MSA sites in early layers, and covariance between MSA sequences is related to Hamming distance.

Finally, to explore the contribution of each column to performance in our regression task, we employed our common logistic model (trained on the means of column attention matrices) to predict Hamming distances using column attentions 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 Supplementary Table 4. 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. Correlations observed between the columns of an MSA can arise both from coevolution due to functional constraints and from phylogeny (see Fig. 5). 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 (Fig. 5b), either independently at equilibrium, or along a phylogenetic tree inferred from the natural MSA using FastTree[33]. The Potts models we used were trained on each of 15 full natural MSAs (see "Methods – Datasets" and Supplementary Table 1) using the generative method bmDCA[26,34]—see "Methods – 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 Supplementary Fig. 4. 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 2, 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 "Methods – 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 resilience might explain 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 (see ref. 28, Sec. 5.1).

Table 2 also shows that plmDCA performs better than MSA Transformer on the synthetic MSAs generated without phylogeny. Because these sequences are sampled independently and at equilibrium from Potts models inferred from the natural MSAs, they are by definition well-described by Potts models. However, these sequences incorporate the imperfections of the inferred Potts models (see the inferred contact maps versus the experimental ones in Supplementary Fig. 2), in addition to lacking the phylogenetic relationships that exist in natural MSAs. These differences with the natural MSAs that were used to train MSA Transformer might explain why it performs less well than plmDCA on these synthetic MSAs, while the opposite holds for natural MSAs (see ref. 28 and Supplementary Figs. 2 and 3). Note that directly comparing the performance of inference between natural and synthetic data is difficult because the ground-truth contacts are not the same and because synthetic data relies on inferred Potts models and inferred phylogenetic trees with their imperfections. However, this does not impair our comparisons of the synthetic datasets generated without and with phylogeny, or of plmDCA and MSA Transformer on the same datasets. Furthermore, an interesting feature that can be observed in Supplementary Fig. 4, and is quantified in Supplementary Table 5, is that MSA Transformer tends to recover the experimental contact maps from our synthetic data generated by bmDCA. Specifically, some secondary structure features that were partially lost in the bmDCA inference and generation process (see the experimental contact maps in Supplementary Fig. 2) 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.

Discussion

MSA Transformer is known to capture structural contacts through its (tied) row attention heads[28]. Here, we showed that it also captures

Table 2 | Impact of phylogeny on contact prediction by plmDCA and MSA Transformer

	ROC-AUC for *N* contacts						ROC-AUC for 2*L* 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

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 "Methods – Synthetic MSA generation via Potts model sampling along inferred phylogenies"). Here, N denotes the number of pairs of residues that have an all-atom minimal distance smaller than 8 Å in the experimental structure in Supplementary Table 1, 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 $\Delta := (A_{\mathrm{eq}} - A_{\mathrm{tree}}) / A_{\mathrm{eq}}$, 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.

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 interleaved row and column 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 phylogenetic relationships 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[35]. For instance, those trained on single protein sequences learn structure and binding sites[36], and those trained on chemical reactions learn how atoms rearrange[37]. 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[38–40] qualitatively captures phylogeny[39]. 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[20,21,41–44], motivating various corrections[17,21,22,24,45–48]. From a theoretical point of view, disentangling these two types of signals is a fundamentally hard problem[49]. 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, our results suggest that the performance of protein language models trained on MSAs could be assessed by evaluating not only how well they capture structural contacts, but also how well they capture phylogenetic relationships. In addition, the ability of protein language models to learn phylogeny could make them particularly well-suited at generating synthetic MSAs capturing the data distribution of natural ones[50]. It also raises the question of their possible usefulness to infer phylogenies and evolutionary histories.

Methods

Datasets

The Pfam database[51] 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[33]. 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

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

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 Supplementary Table 1.

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[52], 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[52]). 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[53], designed to allow the model to distinguish between (a) different aligned positions (columns), and (b) between different sequence positions (rows). (Note that removing the latter kind was shown in ref. 28 to have only limited impact.) 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[54] 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. (Note that both attention blocks, and the feed-forward network, are in fact preceded by layer normalization[55]). The roles of row and column attention in the context of the MLM training objective are illustrated in Fig. 1a. Tied row attention incorporates the expectation that 3D structure should be conserved amongst sequences in an MSA; we refer the reader to ref. 28 for technical details. Column attention works as follows: let $X_j^{(l)}$ be 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 $W_{\text{Q}}^{(l,h)}$, $W_{\text{K}}^{(l,h)}$ and $W_{\text{V}}^{(l,h)}$ (note that these matrices, *mutatis mutandis*, could be of dimension $d \times d'$ with $d' \neq d$), used to obtain three $M \times d$ matrices

$$Q_j^{(l,h)} = X_j^{(l)} W_{\text{Q}}^{(l,h)}, \; K_j^{(l,h)} = X_j^{(l)} W_{\text{K}}^{(l,h)}, \; V_j^{(l,h)} = X_j^{(l)} W_{\text{V}}^{(l,h)}, \tag{2}$$

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

$$A_j^{(l,h)} := \text{softmax}_{\text{row}} \left(\frac{Q_j^{(l,h)} K_j^{(l,h)\text{T}}}{\sqrt{d}} \right), \tag{3}$$

where we denote by $\text{softmax}_{\text{row}}$ the application of $\text{softmax}(\xi_1, \ldots \xi_d) = (e^{\xi_1}, \ldots, e^{\xi_d}) / \sum_{k=1}^{d} e^{\xi_k}$ to each row of a matrix independently, and by $(\cdot)^{\text{T}}$ matrix transposition. 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 $Z_j^{(l,1)} | \cdots | Z_j^{(l,12)}$, 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.

Supervised prediction of Hamming distances

Row i of the column attention matrices $A_j^{(l,h)}$ in Eq. (3) 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 Equation (3) 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 Equation (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

$$\boldsymbol{A}^{(l,h)} := \frac{1}{2(L+1)}\sum_{j=0}^{L}\left(A_j^{(l,h)} + A_j^{(l,h)\mathsf{T}}\right), \text{ with } 1 \le l \le 12 \text{ and } 1 \le h \le 12. \tag{4}$$

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[56] using the `statsmodels` package (version 0.13.2)[57]. 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

$$\mathbb{E}[y\,|\,\boldsymbol{a}] = G_{\beta_0,\boldsymbol{\beta}}(\boldsymbol{a}), \text{ with } G_{\beta_0,\boldsymbol{\beta}}(\boldsymbol{a}) := \sigma\left(\beta_0 + \boldsymbol{a}\boldsymbol{\beta}^{\mathsf{T}}\right), \tag{5}$$

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

$$\ell(\beta_0, \boldsymbol{\beta}\,|\,\boldsymbol{a}, y) = y \log[G_{\beta_0,\boldsymbol{\beta}}(\boldsymbol{a})] + (1 - y)\log[1 - G_{\beta_0,\boldsymbol{\beta}}(\boldsymbol{a})], \tag{6}$$

evaluated over a training set of observations of y and $\boldsymbol{a}$. Note that 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). For simplicity, we refer to these fractional logit models simply as "logistic models". Our general approach to predict Hamming distances is illustrated in Fig. 1b.

Using data from our seed MSAs (cf. Supplementary Table 1), 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[49]—see Fig. 5 for a schematic illustration. 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 "Methods – Datasets"), using FastTree version 2.1[33] 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, version 0.8.12) 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 refs. 26, 34, 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 $x_i \in \{1, \ldots, q\}$ 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. A general Potts model Hamiltonian applied to a sequence $\boldsymbol{x} = (x_1, \ldots, x_L)$ reads

$$H(\boldsymbol{x}) = -\sum_{i=1}^{L} h_i(x_i) - \sum_{j=1}^{L}\sum_{i=1}^{j-1} e_{ij}(x_i, x_j), \tag{7}$$

where the fields $h_i(x_i)$ and couplings $e_{ij}(x_i, x_j)$ are parameters that can be inferred from data by DCA methods[2,20]. In our case, they are inferred from $\mathcal{M}$ by bmDCA[26,34]. The Potts model probability distribution is then given by the Boltzmann distribution associated to the Hamiltonian H in Equation (7):

$$P(\boldsymbol{x}) = \frac{e^{-H(\boldsymbol{x})}}{Z}, \tag{8}$$

where Z is a constant ensuring normalization. In this context, we implement a Metropolis–Hastings algorithm for Markov Chain Monte Carlo (MCMC) sampling from P, where an iteration step consists of a proposed move (mutation) in which a site i is chosen uniformly at random, and its state x_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

$$p = \min[1, \exp(-\Delta H)], \tag{9}$$

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. (7), 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[42], using an open-source implementation available at https://github.com/Bitbol-Lab/Phylogeny-Partners (version 2.0). 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. (9) with the Hamiltonian in Eq. (7). As the length b of a branch gives the estimated number of substitutions that occurred per site along it[33], 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 Supplementary Fig. 1.

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[49], 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. (7)] 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 $(e_{ij}(x, y))_{x,y}$, 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 "Methods – 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` Julia package (https://github.com/pagnani/PlmDCA, version 0.4.1), 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 Supplementary Table 1—see Supplementary Fig. 2. 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 $M_{\mathrm{sub}} \approx 380$ for family PF13354 due to its greater length, and $M_{\mathrm{sub}} \approx 500$ for all other families. Then, we computed for each subsample a matrix of coupling scores using MSA Transformer's row attention heads and the estimated contact probabilities from the aforementioned logistic regression. Finally, we averaged the resulting 10 matrices to obtain a single matrix of coupling scores. We used a similar strategy (and the same randomly sampled row indices) to infer structural contact scores from the natural MSAs—see Supplementary Fig. 3. Consistently with findings in ref. 28, MSA Transformer generally performs better than plmDCA (Supplementary Fig. 2) at contact inference.

Reporting summary

Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

All sequence data used or generated in our work has been deposited in https://zenodo.org/record/7096792. We made use of the following PDB structures: 4D81 [https://doi.org/10.2210/pdb4D81/pdb], 1L7V [https://doi.org/10.2210/pdb1L7V/pdb], 3UP1 [https://doi.org/10.2210/pdb3UP1/pdb], 3ILH [https://doi.org/10.2210/pdb3ILH/pdb], 3NNH [https://doi.org/10.2210/pdb3NNH/pdb], 4R2A [https://doi.org/10.2210/pdb4R2A/pdb], 1OCK [https://doi.org/10.2210/pdb1OCK/pdb], 3EX7 [https://doi.org/10.2210/pdb3EX7/pdb], 4REX [https://doi.org/10.2210/pdb4REX/pdb], 3DGE [https://doi.org/10.2210/pdb3DGE/pdb], 1BE9 [https://doi.org/10.2210/pdb1BE9/pdb], 4M57 [https://doi.org/10.2210/pdb4M57/pdb], 3G7E [https://doi.org/10.2210/pdb3G7E/pdb], 1FHG [https://doi.org/10.2210/pdb1FHG/pdb], 6QW8 [https://doi.org/10.2210/pdb6QW8/pdb].

Code availability

Our code is available at https://zenodo.org/record/7096792.

References

1. de Juan, D., Pazos, F. & Valencia, A. Emerging methods in protein co-evolution. *Nat. Rev. Genet.* **14**, 249–261 (2013).
2. Cocco, S., Feinauer, C., Figliuzzi, M., Monasson, R. & Weigt, M. Inverse statistical physics of protein sequences: a key issues review. *Rep. Prog. Phys.* **81**, 032601 (2018).
3. Bahdanau, D., Cho, K. & Bengio, Y. Neural machine translation by jointly learning to align and translate (ICLR 2015). *arXiv* https://doi.org/10.48550/arXiv.1409.0473 (2014).
4. Vaswani, A. et al. Attention is all you need. *Adv. Neural Inf. Process. Syst.* **30**, 5998–6008 (2017).
5. Elnaggar, A. et al. ProtTrans: towards cracking the language of life's code through self-supervised learning. *bioRxiv* https://doi.org/10.1101/2020.07.12.199554 (2020).
6. Rives, A. et al. Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *Proc. Natl. Acad. Sci. USA* **118** https://www.pnas.org/content/118/15/e2016239118 (2021).
7. Rao, R., Meier, J., Sercu, T., Ovchinnikov, S. & Rives, A. Transformer protein language models are unsupervised structure learners. In *International Conference on Learning Representations* https://openreview.net/forum?id=fylclEqgvgd (2021).
8. Choromanski, K. et al. Rethinking attention with Performers. In *International Conference on Learning Representations*. https://openreview.net/forum?id=Ua6zuk0WRH (2021).
9. Madani, A. et al. ProGen: Language modeling for protein generation. *bioRxiv* https://doi.org/10.1101/2020.03.07.982272 (2020).
10. Madani, A. et al. Deep neural language modeling enables functional protein generation across families. *bioRxiv* https://doi.org/10.1101/2021.07.18.452833 (2021).
11. Bhattacharya, N. et al. Interpreting potts and transformer protein models through the lens of simplified attention. *Pac. Symp. Biocomput.* **27**, 34–45 (2022).
12. Jumper, J. et al. Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583–589 (2021).
13. Baek, M. et al. Accurate prediction of protein structures and interactions using a three-track neural network. *Science* **373**, 871–876 (2021).
14. Chowdhury, R. et al. Single-sequence protein structure prediction using language models from deep learning. *bioRxiv* https://doi.org/10.1101/2021.08.02.454840 (2021).
15. Casari, G., Sander, C. & Valencia, A. A method to predict functional residues in proteins. *Nat. Struct. Biol.* **2**, 171–178 (1995).
16. Socolich, M. et al. Evolutionary information for specifying a protein fold. *Nature* **437**, 512–518 (2005).
17. Dunn, S. D., Wahl, L. M. & Gloor, G. B. Mutual information without the influence of phylogeny or entropy dramatically improves residue contact prediction. *Bioinformatics* **24**, 333–340 (2008).
18. Halabi, N., Rivoire, O., Leibler, S. & Ranganathan, R. Protein sectors: evolutionary units of three-dimensional structure. *Cell* **138**, 774–786 (2009).
19. Lapedes, A. S., Giraud, B. G., Liu, L. & Stormo, G. D. Correlated mutations in models of protein sequences: phylogenetic and structural effects. In *Statistics in molecular biology and genetics – IMS Lecture Notes – Monograph Series*, vol. 33, 236–256 (Institute of Mathematical Statistics, 1999). https://doi.org/10.1214/lnms/1215455556.
20. Weigt, M., White, R. A., Szurmant, H., Hoch, J. A. & Hwa, T. Identification of direct residue contacts in protein-protein interaction by message passing. *Proc. Natl. Acad. Sci. USA* **106**, 67–72 (2009).
21. Marks, D. S. et al. Protein 3D structure computed from evolutionary sequence variation. *PLoS ONE* **6**, e28766 (2011).
22. Morcos, F. et al. Direct-coupling analysis of residue coevolution captures native contacts across many protein families. *Proc. Natl. Acad. Sci. USA* **108**, E1293–1301 (2011).
23. Sułkowska, J. I., Morcos, F., Weigt, M., Hwa, T. & Onuchic, J. N. Genomics-aided structure prediction. *Proc. Natl. Acad. Sci. USA* **109**, 10340–10345 (2012).
24. Ekeberg, M., Lovkvist, C., Lan, Y., Weigt, M. & Aurell, E. Improved contact prediction in proteins: using pseudolikelihoods to infer Potts models. *Phys. Rev. E* **87**, 012707 (2013).
25. Ekeberg, M., Hartonen, T. & Aurell, E. Fast pseudolikelihood maximization for direct-coupling analysis of protein structure from many homologous amino-acid sequences. *J. Comput. Phys.* **276**, 341–356 (2014).
26. Figliuzzi, M., Barrat-Charlaix, P. & Weigt, M. How pairwise coevolutionary models capture the collective residue variability in proteins? *Mol. Biol. Evol.* **35**, 1018–1027 (2018).
27. Abriata, L. A., Tamó, G. E., Monastyrskyy, B., Kryshtafovych, A. & Dal Peraro, M. Assessment of hard target modeling in CASP12 reveals an emerging role of alignment-based contact prediction methods. *Proteins* **86**, 97–112 (2018).
28. Rao, R. M. et al. MSA Transformer. *Proceedings of the 38th International Conference on Machine Learning* **139**, 8844–8856 (2021).
29. Meier, J. et al. Language models enable zero-shot prediction of the effects of mutations on protein function. In Beygelzimer, A., Dauphin, Y., Liang, P. & Vaughan, J. W. (eds.) *Advances in Neural Information Processing Systems* (2021). https://openreview.net/forum?id=uXc42E9ZPFs.
30. Hie, B. L., Yang, K. K. & Kim, P. S. Evolutionary velocity with protein language models predicts evolutionary dynamics of diverse proteins. *Cell Systems* **13**, 274–285.e6 (2022).
31. Hawkins-Hooker, A., Jones, D. T. & Paige, B. MSA-conditioned generative protein language models for fitness landscape modelling and design. In *Machine Learning for Structural Biology Workshop, NeurIPS* (2021). https://www.mlsb.io/papers_2021/MLSB2021_MSA-Conditioned_Generative_Protein_Language.pdf.
32. Gordon, Y. On Milman's inequality and random subspaces which escape through a mesh in $\mathbb{R}^n$. In Lindenstrauss, J. & Milman, V. D. (eds.) *Geometric Aspects of Functional Analysis*, 84–106 (Springer, Berlin, Heidelberg, 1988). https://doi.org/10.1007/BFb0081737.
33. Price, M. N., Dehal, P. S. & Arkin, A. P. FastTree 2 - approximately maximum-likelihood trees for large alignments. *PLOS ONE* **5**, 1–10 (2010).
34. Russ, W. P. et al. An evolution-based model for designing chorismate mutase enzymes. *Science* **369**, 440–445 (2020).
35. Rogers, A., Kovaleva, O. & Rumshisky, A. A primer in BERTology: what we know about how BERT works. *Transactions of the Association for Computational Linguistics* **8**, 842–866 (2020).

36. Vig, J. et al. BERTology meets biology: Interpreting attention in protein language models. In *International Conference on Learning Representations* (2021). https://openreview.net/forum?id=YWtLZvLmud7.
37. Schwaller, P., Hoover, B., Reymond, J. L., Strobelt, H. & Laino, T. Extraction of organic chemistry grammar from unsupervised learning of chemical reactions. Sci. Adv. **7**, https://doi.org/10.1126/sciadv.abe4166 (2021).
38. Riesselman, A. J., Ingraham, J. B. & Marks, D. S. Deep generative models of genetic variation capture the effects of mutations. *Nat. Methods* **15**, 816–822 (2018).
39. Ding, X., Zou, Z. & Brooks III, C. L. Deciphering protein evolution and fitness landscapes with latent space models. *Nat. Commun.* **10**, 5644 (2019).
40. McGee, F. et al. The generative capacity of probabilistic protein sequence models. *Nat. Commun.* **12**, 6302 (2021).
41. Qin, C. & Colwell, L. J. Power law tails in phylogenetic systems. *Proc. Natl. Acad. Sci. USA* **115**, 690–695 (2018).
42. Vorberg, S., Seemayer, S. & Söding, J. Synthetic protein alignments by CCMgen quantify noise in residue-residue contact prediction. *PLoS Comput. Biol.* **14**, 1–25 (2018).
43. Rodriguez Horta, E., Barrat-Charlaix, P. & Weigt, M. Toward inferring Potts models for phylogenetically correlated sequence data. *Entropy* **21** https://www.mdpi.com/1099-4300/21/11/1090 (2019).
44. Rodriguez Horta, E. & Weigt, M. On the effect of phylogenetic correlations in coevolution-based contact prediction in proteins. *PLoS Comput. Biol.* **17** https://doi.org/10.1371/journal.pcbi.1008957 (2021).
45. Lichtarge, O., Bourne, H. R. & Cohen, F. E. An evolutionary trace method defines binding surfaces common to protein families. *J. Mol. Biol.* **257**, 342–358 (1996).
46. Hockenberry, A. J. & Wilke, C. O. Phylogenetic weighting does little to improve the accuracy of evolutionary coupling analyses. *Entropy* **21**, https://doi.org/10.3390/e21101000 (2019).
47. Malinverni, D. & Barducci, A. Coevolutionary analysis of protein subfamilies by sequence reweighting. *Entropy* **21**, 1127 (2020).
48. Colavin, A., Atolia, E., Bitbol, A.-F. & Huang, K. C. Extracting phylogenetic dimensions of coevolution reveals hidden functional signals. *Sci. Rep.* **12**, 820 (2022).
49. Weinstein, E. N., Amin, A. N., Frazer, J. & Marks, D. S. Non-identifiability and the blessings of misspecification in models of molecular fitness and phylogeny. *bioRxiv* https://doi.org/10.1101/2022.01.29.478324 (2022).
50. Sgarbossa, D., Lupo, U. & Bitbol, A.-F. Generative power of a protein language model trained on multiple sequence alignments. *bioRxiv* https://doi.org/10.1101/2022.04.14.488405 (2022).
51. Mistry, J. et al. Pfam: The protein families database in 2021. *Nucleic Acids Res.* **49**, D412–D419 (2020).
52. Devlin, J., Chang, M.-W., Lee, K. & Toutanova, K. 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)*, 4171-4186 (Association for Computational Linguistics, Minneapolis, Minnesota, 2019). https://aclanthology.org/N19-1423.
53. Gehring, J., Auli, M., Grangier, D., Yarats, D. & Dauphin, Y. N. Convolutional sequence to sequence learning. In Precup, D. & Teh, Y. W. (eds.) *Proceedings of the 34th International Conference on Machine Learning*, vol. 70 of *Proceedings of Machine Learning Research*, 1243–1252 (PMLR, 2017). https://proceedings.mlr.press/v70/gehring17a.html.
54. Ho, J., Kalchbrenner, N., Weissenborn, D. & Salimans, T. Axial attention in multidimensional transformers. *arXiv* https://doi.org/10.48550/arXiv.1912.12180 (2019).
55. Ba, J. L., Kiros, J. R. & Hinton, G. E. Layer normalization. *arXiv* https://doi.org/10.48550/arXiv.1607.06450 (2016).
56. Papke, L. E. & Wooldridge, J. M. Econometric methods for fractional response variables with an application to 401(k) plan participation rates. *J. Appl. Econ.* **11**, 619–632 (1996).
57. Seabold, S. & Perktold, J. Statsmodels: Econometric and statistical modeling with Python. In *9th Python in Science Conference* (2010). https://doi.org/10.25080/Majora-92bf1922-011.

Acknowledgements

The authors thank Mohammed AlQuraishi for inspiring discussions, Sergey Ovchinnikov for valuable feedback on the first version of this manuscript, Tom Sercu for clarifying some aspects of MSA Transformer's pre-training, and Alexander Rives for interesting conversations. Code by Andonis Gerardos was used to generate synthetic MSAs along prescribed phylogenies. 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.).

Author contributions

All authors participated in the design of the project. U.L. and D.S. wrote the software. U.L. built the datasets and performed the bmDCA analysis. U.L. and D.S. performed the analysis of MSA Transformer column attention. All authors interpreted the results. A.-F.B. supervised the project. U.L. and A.-F.B. wrote the manuscript. All authors reviewed and edited the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41467-022-34032-y.

Correspondence and requests for materials should be addressed to Umberto Lupo or Anne-Florence Bitbol.

Peer review information *Nature Communications* thanks Arne Elofsson and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

Reprints and permission information is available at http://www.nature.com/reprints

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.