Protein language models trained on multiple sequence alignments learn phylogenetic relationships

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

Abstract

Self-supervised neural language models (LMs) with attention have recently been applied to biological sequence data, advancing structure, function and mutational effect prediction. Some protein LMs, including *MSA Transformer* and *AlphaFold*'s "EvoFormer", take multiple sequence alignments (MSAs) as inputs. We demonstrate that simple, and universal, combinations of MSA Transformer's column attentions strongly correlate with Hamming distances between sequences in MSAs. Therefore, MSA-based LMs encode detailed phylogenetic relationships. We further show that these models can separate coevolutionary signals encoding functional and structural constraints from correlations arising from 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 Potts models.

1. Introduction

Recently, self-supervised deep learning models based on natural language processing methods, especially attention and transformers, 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 (Elnaggar et al., 2020; Rives et al., 2021; Rao et al., 2021a; Choromanski et al., 2021; Madani et al., 2020; 2021). These models, which capture long-range dependencies, learn rich representations of protein sequences, and can be employed for multiple tasks (Meier et al.; Hawkins-Hooker et al., 2021; Hie et al., 2022). Neural network architectures based on attention are also employed in *AlphaFold* (Jumper et al., 2021), *RoseTTAFold* (Baek et al., 2021) and *RGN2* (Chowdhury et al., 2021), and thus 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. 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* (Rao et al., 2021b) and the "Evoformer" blocks in AlphaFold (Jumper et al., 2021), 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 (Rao et al., 2021b), outperforming language models trained on individual sequences, as well as Potts models, also known in the field as DCA (Direct Coupling Analysis) (Cocco et al., 2018).

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? Is MSA Transformer able to separate coevolutionary signals encoding functional and structural constraints from phylogenetic correlations arising from historical contingency?

Datasets and code for reproducing our analyses can be found at `https://github.com/Bitbol-Lab/Phylogeny-MSA-Transformer`, and further details can be found in (Lupo et al., 2022).

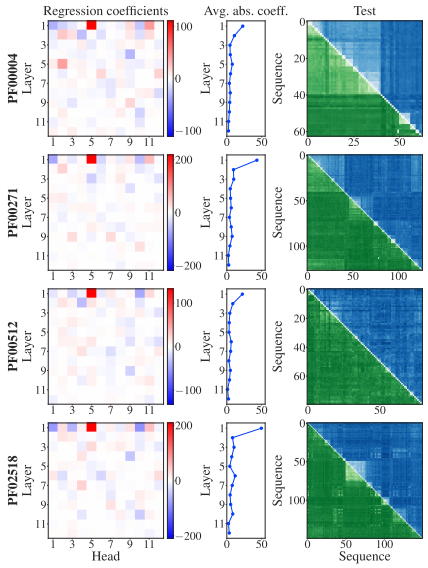
[1]Institute of Bioengineering, School of Life Sciences, École Polytechnique Fédérale de Lausanne (EPFL), CH-1015 Lausanne, Switzerland [2]SIB Swiss Institute of Bioinformatics, CH-1015 Lausanne, Switzerland. Correspondence to: Umberto Lupo <umberto.lupo@epfl.ch>, Anne-Florence Bitbol <anne-florence.bitbol@epfl.ch>.

The 2022 ICML Workshop on Computational Biology.

2. Results

2.1. Column attention heads capture Hamming distances in separate MSAs

We first considered separately each of 15 different Pfam seed MSAs (see Appendix A.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 based on the column-wise means of MSA Transformer's column attention heads on all pairwise Hamming distances in the training set – see Appendix A.2 for details. 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 Figure 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 Figure 1(a), third column, for results on four example MSAs. 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 Figure 1(b). A striking result from our analysis is that the regression coefficients appear to be similar across MSAs – see Figure 1(a), first column. To quantify this, we computed the Pearson correlations between the regression coefficients learnt on "larger" seed MSAs (i.e. on the 7 MSAs with ≥ 100 sequences and ≥ 30 residues). The values of these correlations are between 0.69 and 0.87 (mean: 0.80), demonstrating 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.** Each MSA is randomly split into a training set comprising 70% of its sequences and a test set composed of the remaining sequences. (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 the test sets from four example MSAs (third column), ground truth Hamming distances are shown in the upper triangle (blue) and predicted Hamming distances in the lower triangle (green). 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.

2.2. 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 2 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 Table 1, 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.

(Rao et al., 2021b) showed that some column attention matrices, summed along one of their dimensions, correlate with phylogenetic sequence weights (see Appendix A.1). This indicates that the model is, in part, attending to maximally diverse sequences. Our study demonstrates that MSA Trans-

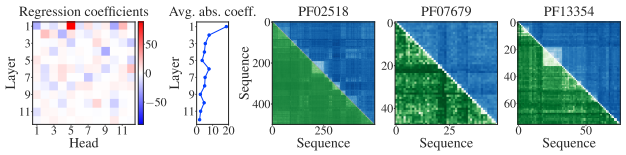


Figure 2. **Fitting a single logistic model to predict Hamming distances.** Regression coefficients (first panel) and their absolute values averaged over heads for each layer (second panel) are shown. For the three test MSAs, ground truth Hamming distances are shown in the upper triangle (blue) and predicted Hamming distances in the lower triangle (green) (last three panels).

Table 1. **Quality of fit for our logistic model trained on Hamming distances and column attentions from several MSAs.** For the logistic model described in Section 2.2, 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	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**

former 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 Figure 2 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 Figures 1(a) and 2 with Fig. 5 in Ref. (Rao et al., 2021b) shows that different attention heads are important in our study versus in the analysis of Ref. (Rao et al., 2021b).

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.[1] The results obtained in this case for our regression task are reported in Table A2. They demonstrate that, although random initialization can yield better performance than random guessing,[2] the trained MSA Transformer gives vastly superior results.

For each layer and attention head in the network, MSA Transformer computes one matrix of column attention values per site – see Equation (1). This is in contrast with row attention, which is tied. 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.[3]

2.3. MSA Transformer efficiently disentangles correlations from contacts and phylogeny

MSA Transformer is known to capture three-dimensional contacts through its (tied) row attention heads (Rao et al., 2021b), 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` (Price et al., 2010). The Potts models we used were trained on each of 15 "full" natural MSAs (see Appendix A.1) using the generative method bmDCA (Figliuzzi et al., 2018) – see Appendix A.3. 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".

Examples of contact maps inferred by plmDCA (Ekeberg et al., 2013) and by MSA Transformer for our synthetic

[1] We thank Tom Sercu for sharing details about MSA Transformer's initialization with us.

[2] Some intuition for this result is provided by the Johnson–Lindenstrauss Lemma and Gordon's Theorem (Gordon, 1988).

[3] We thank Sergey Ovchinnikov for useful discussions on these points.

datasets are shown in Figure A1. 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. For both plmDCA and MSA Transformer, we quantified the degradation in performance caused by the injection of phylogeny by computing the relative drop, Δ, in the area under the receiver operating characteristic curve (ROC-AUC) for contact prediction. Results are reported in Table 2, for all Pfam families and for two different cutoffs on the number of contacts. 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 Appendix A.3). The conclusion is the same if 10 (or 6, for Pfam family PF13354) twice-deeper subsampled MSAs are employed.

Table 2. **Impact of phylogeny on contact prediction by plmDCA and MSA Transformer.** Ground truth contacts are defined as the N or $2L$ pairs with top coupling scores, where L the length of the MSA and N denotes the number of pairs of residues that have an all-atom distance smaller than 8 Å in the experimental structure in Table A1, excluding pairs at positions i, j with $|i - j| \leq 4$.

	Δ for N contacts		Δ for $2L$ contacts	
Pfam ID	plmDCA	MSA Tr.	plmDCA	MSA Tr.
PF00004	0.33	0.04	0.34	0.11
PF00005	0.28	0.03	0.23	−0.01
PF00041	0.25	0.10	0.22	0.09
PF00072	0.23	0.10	0.14	0.08
PF00076	0.25	0.05	0.25	0.05
PF00096	0.39	0.21	0.41	0.30
PF00153	0.26	0.24	0.21	0.28
PF00271	0.32	0.07	0.29	0.10
PF00397	0.33	0.15	0.34	0.22
PF00512	0.21	0.08	0.20	0.08
PF00595	0.33	0.14	0.33	0.18
PF01535	0.23	0.05	0.18	0.01
PF02518	0.27	0.09	0.20	0.12
PF07679	0.26	0.05	0.19	0.05
PF13354	0.18	0.14	0.21	0.19
Average	0.27	0.10	0.25	0.12

These results demonstrate that contact inference by MSA Transformer is less deteriorated by phylogenetic correlations than contact inference by Potts models. 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 Sec. 5.1 in Ref. (Rao et al., 2021b)).

3. Discussion

MSA Transformer is known to capture structural contacts through its (tied) row attention heads (Rao et al., 2021b). Here, we showed that it also captures Hamming distances, and thus phylogenetic information, through its 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. 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.

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. 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. Phylogenetic correlations are known to obscure the identification of structural contacts by traditional coevolution methods, in particular by inferred Potts models, motivating various corrections. From a theoretical point of view, disentangling these two types of signals is a fundamentally hard problem (Weinstein et al., 2022). 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 models, 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. More generally, the ability of protein language models to learn phylogeny raises the question of their possible usefulness to infer phylogenies and evolutionary histories.

References

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

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

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

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

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

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

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

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

Gordon, Y. On Milman's inequality and random subspaces which escape through a mesh in $\mathbb{R}^n$. In Lindenstrauss, J. and Milman, V. D. (eds.), *Geometric Aspects of Functional Analysis*, pp. 84–106. Springer, Berlin, Heidelberg, 1988. ISBN 978-3-540-39235-4.

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

Hie, B. L., Yang, K. K., and Kim, P. S. Evolutionary velocity with protein language models predicts evolutionary dynamics of diverse proteins. *Cell Systems*, 13(4):274–285.e6, 2022.

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

Lupo, U., Sgarbossa, D., and Bitbol, A.-F. Protein language models trained on multiple sequence alignments learn phylogenetic relationships. *bioRxiv*, 10.1101/2022.03.29.486219, 2022.

Madani, A., McCann, B., Naik, N., Keskar, N. S., Anand, N., Eguchi, R. R., Huang, P.-S., and Socher, R. ProGen: Language modeling for protein generation, 2020.

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

Meier, J., Rao, R., Verkuil, R., Liu, J., Sercu, T., and Rives, A. Language models enable zero-shot prediction of the effects of mutations on protein function. In *Advances in Neural Information Processing Systems*.

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

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

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

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

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

A. Materials and methods

A.1. Datasets

For each of its families, the Pfam database[4] provides an expert-curated "seed" alignment that contains a representative set of sequences, as well as deeper "full" alignments that are automatically built by searching against a large sequence database using a profile hidden Markov model built from the seed alignments. 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*. For each family, we retrieved one experimental three-dimensional protein structure from the PDB (`https://www.rcsb.org`). 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. 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 `https://github.com/matteofigliuzzi/bmDCA` (publication date: Dec. 2017) and were previously considered in Ref. (Figliuzzi et al., 2018). We computed the *effective depth* of each MSA as $M_{\text{eff}}^{(\delta)} := \sum_{i=1}^{M} w_i$, with $w_i := |\{i' : d_{\text{H}}(\boldsymbol{x}^{(i)}, \boldsymbol{x}^{(i')}) < \delta\}|^{-1}$, where M is the depth of the MSA, $\boldsymbol{x}^{(i)}$ is the i-th sequence in the MSA, $d_{\text{H}}(\boldsymbol{x}, \boldsymbol{y})$ 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. Information about our MSAs is summarized in Table A1.

Table A1. **Pfam families and MSAs used in this work.** L denotes the length of an MSA, M its depth, and $M_{\text{eff}}^{(0.2)}$ its effective depth.

		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

A.2. Supervised prediction of Hamming distances

We used the pre-trained MSA Transformer model introduced in Ref. (Rao et al., 2021b), retrieved from the Python Package Index as `fair-esm 0.4.0`. We recall that this model comprises 12 successive layers of 12 axial attention blocks. The axial attention blocks are executed in parallel within each layer, and consist of a tied row attention block, followed by a column attention block and, finally, by a feed-forward network – see Ref. (Rao et al., 2021b) for details. When receiving an MSA with L columns and M rows as input, the model computes, for each layer $1 \leq l \leq 12$, for each head $1 \leq h \leq 12$, and for each MSA column j (plus an additional "beginning-of-sentence" position corresponding to $j = 0$), a $M \times M$ column attention matrix $A_j^{(l,h)}$. We predict the Hamming distance y between the i-th and the i'-th sequence, in an MSA $\mathcal{M}$ of length L, using the entries $a_{i,i'}^{(l,h)}$ at position (i, i') (henceforth $a^{(l,h)}$ for brevity) from the 144 matrices

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

[4] `https://pfam.xfam.org/`

We fit fractional logit models via quasi-maximum likelihood estimation using the `statsmodels` package.[5] 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 $\mathbb{E}[y \,|\, \boldsymbol{a}] = G_{\beta_0, \boldsymbol{\beta}}(\boldsymbol{a})$, with $G_{\beta_0, \boldsymbol{\beta}}(\boldsymbol{a}) := \sigma(\beta_0 + \boldsymbol{a}\boldsymbol{\beta}^{\mathrm{T}})$, 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 $\ell(\beta_0, \boldsymbol{\beta} \,|\, \boldsymbol{a}, y) := y \log[G_{\beta_0, \boldsymbol{\beta}}(\boldsymbol{a})] + (1 - y) \log[1 - G_{\beta_0, \boldsymbol{\beta}}(\boldsymbol{a})]$, evaluated over a training set of observations of y and $\boldsymbol{a}$. We refer to these models simply as "logistic models".

Using data from our seed MSAs (cf. Table A1), 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 $I_{\mathcal{M},\,\mathrm{train}}$ and $I_{\mathcal{M},\,\mathrm{test}}$, with $I_{\mathcal{M},\,\mathrm{train}}$ 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 $I_{\mathcal{M},\,\mathrm{train}}$, and as test data the Hamming distances, and column attentions, coming from pairs of indices in $I_{\mathcal{M},\,\mathrm{test}}$. 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.

A.3. Synthetic MSA generation via Potts model sampling along inferred phylogenies

We inferred unrooted phylogenetic trees from our full MSAs (see Appendix A.1), using `FastTree 2.1` (Price et al., 2010) with its default settings.[6] Then, we fitted Potts models on each of these MSAs using bmDCA (Figliuzzi et al., 2018) (`https://github.com/ranganathanlab/bmDCA`) with its default hyperparameters. The choice of bmDCA is motivated by the fact that, as has been shown in Refs. (Figliuzzi et al., 2018), 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, e.g. pseudo-likelihood DCA (plmDCA) (Ekeberg et al., 2013), which is the DCA standard for contact prediction but cannot faithfully reproduce empirical one- and two-body marginals.

Let $\mathcal{M}$ denote an arbitrary MSA from our set of full MSAs, L its length, and M its depth. Using the phylogenetic tree and Potts model inferred by bmDCA from $\mathcal{M}$, we generated a synthetic MSA without phylogeny by equilibrium Markov Chain Monte Carlo (MCMC) sampling from the inferred Potts model, using a Metropolis–Hastings algorithm in which each step consists of a proposed move ("mutation") in which a site i in a sequence of L amino-acid is chosen uniformly at random, and its state may be changed into another state chosen uniformly at random. We started from a set of M randomly and independently initialized sequences, and proposed a total number N of mutations on each sequence independently. 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 (Figliuzzi et al., 2018). We thus used the value of N estimated by bmDCA at the end of training.

We also generated synthetic data using MCMC sampling along our inferred phylogenetic trees. We started from an equilibrium ancestor sequence sampled as explained above, and placed it at an arbitrary root (root placement does not matter; see below). Then, we evolved this sequence by successive duplication (at each branching of the tree) and mutation events (along each branch). Mutations were modeled using for acceptance the standard Metropolis criterion given the inferred Potts model Hamiltonian. As the length b of a branch gives the estimated number of substitutions that occurred per site along it (Price et al., 2010), we generate data by making a number of accepted mutations on this branch equal to the integer closest to bL. Our procedure for generating MSAs along a phylogeny is independent of the placement of the tree's root. Indeed, 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.

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 is to (1) learn fields and couplings 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) (Dunn et al., 2008), yielding a coupling score E_{ij}. Top scoring pairs of sites are then predicted as being contacts. In the case of MSA Transformer (Rao et al., 2021b), a single logistic

[5] `https://www.statsmodels.org`

[6] 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 our analysis, which relies on fitting Potts models accurately.

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. We applied these inference techniques, normally used to predict structural contacts, on our *synthetic* MSAs generated without and with phylogeny. As proxies for structural contacts, we used the pairs of sites with top coupling scores in the Potts models used to generate the MSAs. As a DCA method to infer these coupling scores, we used plmDCA (Ekeberg et al., 2013) 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 (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 A1).

While Potts models need to be fitted on deep MSAs to achieve good contact prediction, MSA Transformer's memory requirements are too large for us to run it on any of our synthetic MSAs in its 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 estimated contact probabilities from the aforementioned logistic regression. Finally, we averaged the resulting 10 matrices to obtain a single matrix of coupling scores.

B. Common logistic model using MSA Transformer with random weights

Table A2. We reinitialized MSA Transformer's parameters to random values, using the protocols originally used in pre-training (see Section 2.2). Results for the same task as in Table 1 are shown.

Family	R^2	Pearson	Slope
PF00004	0.31	0.67	1.60
PF00005	0.33	0.59	0.96
PF00041	0.02	0.43	1.02
PF00072	0.45	0.67	1.07
PF00076	0.30	0.56	1.10
PF00096	-0.13	0.24	0.52
PF00153	0.07	0.32	0.66
PF00271	0.13	0.46	1.25
PF00397	-0.31	0.39	1.19
PF00512	0.02	0.35	0.73
PF00595	0.49	0.70	1.09
PF01535	-0.17	0.20	0.63
PF02518	**-0.09**	**0.32**	**1.07**
PF07679	**0.29**	**0.57**	**0.91**
PF13354	**0.35**	**0.62**	**1.15**

C. Contact prediction on synthetic MSAs

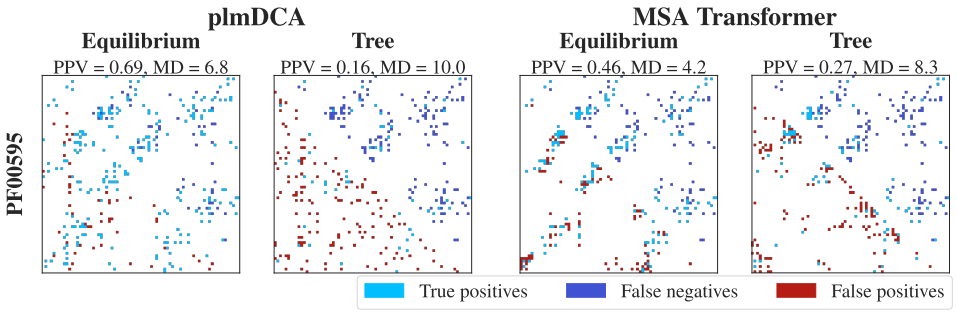


Figure A1. **Predicted contact maps using plmDCA and MSA Transformer on synthetic MSAs, versus ground truth "contact maps" defined by top Potts model couplings, for Pfam family PF00595.** "Contact maps" containing $2L$ contacts, and obtained from the ground-truth couplings in the bmDCA model used to generate the synthetic MSAs either without phylogeny ("Equilibrium") or with phylogeny ("Tree") – see Appendix A.3 – 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 in each case. 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 A1).