Direct coupling analysis and the attention mechanism

BMC Bioinformatics

RESEARCH

Open Access

Direct coupling analysis and the attention mechanism

Francesco Caredda[1*†] and Andrea Pagnani[1,2,3†]

†Francesco Caredda and Andrea Pagnani have authors contributed equally to this work.

*Correspondence: francesco.caredda@polito.it

[1] DISAT, Politecnico di Torino, Corso Duca degli Abruzzi, I-10129 Torino, Italy
[2] Italian Institute for Genomic Medicine, IRCCS Candiolo, SP-142, I-10060 Candiolo, Italy
[3] INFN, Sezione di Torino, Via Pietro Giuria, I-10125 Torino, Italy

Abstract

Proteins are involved in nearly all cellular functions, encompassing roles in transport, signaling, enzymatic activity, and more. Their functionalities crucially depend on their complex three-dimensional arrangement. For this reason, being able to predict their structure from the amino acid sequence has been and still is a phenomenal computational challenge that the introduction of *AlphaFold* solved with unprecedented accuracy. However, the inherent complexity of *AlphaFold*'s architectures makes it challenging to understand the rules that ultimately shape the protein's predicted structure. This study investigates a single-layer unsupervised model based on the attention mechanism. More precisely, we explore a Direct Coupling Analysis (DCA) method that mimics the attention mechanism of several popular *Transformer* architectures, such as *AlphaFold* itself. The model's parameters, notably fewer than those in standard DCA-based algorithms, can be directly used for extracting structural determinants such as the contact map of the protein family under study. Additionally, the functional form of the energy function of the model enables us to deploy a multi-family learning strategy, allowing us to effectively integrate information across multiple protein families, whereas standard DCA algorithms are typically limited to single protein families. Finally, we implemented a generative version of the model using an autoregressive architecture, capable of efficiently generating new proteins in silico.

Keywords: Protein structure prediction, Attention mechanism, Direct coupling analysis, Transformer

Introduction

Proteins constitute a diverse category of biological compounds constructed from a set of 20 amino acids. Within an organism, they serve various functions, including structural support, mobility, and enzymatic activities. The effectiveness of a protein is intricately linked to its three-dimensional arrangement, known as its tertiary structure. This structure dictates the protein's biological functionality when isolated and its interactions with other molecules within the cellular environment. Under physiological conditions, a protein's three-dimensional configuration is uniquely determined by its amino acid sequence [1]. Understanding this dependence is theoretically and computationally challenging due to the system's complexity. Only recently, the problem of predicting the fold of an amino acid sequence has seen a historic computational breakthrough

thanks to *AlphaFold* in 2021 [2], which led its authors to receive the 2024 Nobel Prize in Chemistry.

AlphaFold exploits decades of research in computational biology and recent developments in machine learning. The foundational idea, which has been the center of this research area for years, is that evolutionary information can be extracted and used to determine patterns in phylogenetically related protein sequences (viz. homologs). Indeed, in the course of evolution, their structure must be conserved to preserve the functionality of a class of proteins. Natural selection imposes constraints on single active sites or multi-amino acid motifs that are fundamental for the correct sequence folding. This leads to the idea of conservation and co-evolution [3, 4]. Given a Multiple Sequence Alignment (MSA) [5], single- and pair-wise frequencies of amino acids along different positions are enough to extract summary statistics that can be used to determine structural information by inferring the parameters of a Potts model, in what is commonly known as Direct Coupling Analysis (DCA) [6, 7]. The inference has been implemented in various ways and with various degrees of approximation during the last decade: mean-field DCA [7, 8], methods based on Gaussian approximations [9], pseudo-likelihood-based methods (PlmDCA, [10]). Although the original application was the prediction of protein residue contacts, other exciting applications have emerged more recently: inference of protein-protein interactions [11, 12], conformational plasticity [13, 14], inference of mutational landscapes [15–18] and in silico generation of sequences representative of the full statistics of the original protein family (bmDCA [19, 20], ArDCA, [21]).

A similar strategy to DCA methods is adopted by *AlphaFold* [2], where the self-attention mechanism allows for a direct representation of correlations over the MSA, albeit through multiple layers [22, 23]. The attention mechanism was originally introduced in the context of Natural Language Processing (NLP) to overcome the limitations of sequential encoder-decoder architectures [24]. The basic idea is that long-range correlations within a dataset can be captured by a so-called *attention map*, encoding a custom functional relation between features of the dataset. For instance, in NLP the correlations to capture emerge at the semantic level [24], whereas in the case of structural protein inference from homology modeling, at the level of the individual residues as in the *Evoformer* architectural block in *AlphaFold*, in which a contact representation of the residues in a sequence is updated by conservation and co-evolution information extracted from an MSA and processed through self-attention layers.

Background and aims

We analyze the *factored attention layer* defined by Bhattacharya et al. [25] as a simplified version of the dot product self-attention mechanism [22]. In a factored attention layer, the positional degrees of freedom of the amino acid sequence are decoupled from the *color/amino acid* degrees of freedom representing each possible amino acid. This factorization separates the signal from a specific protein family and the signal due to the nature of amino acid interaction shared across multiple families. In their work, they demonstrate that a factored attention model can be traced back to a generalized Potts model [26], and ultimately all the methods developed in the context of DCA [7] can be used to obtain a contact prediction algorithm.

Interestingly, in [25] is shown that even if the number of parameters of the *factored attention* is significantly lower compared to the standard DCA models, their performance is almost equivalent. However, they limit their analysis to a contact score obtained in the same manner as for a generic Potts model, while here we push forward this analysis by studying the contact prediction obtained directly from the attention matrices of a factored attention model. More precisely, we show that the accuracy of the Frobenius and Attention scores is compatible across multiple protein families. We argue that this is a more direct way to understand the inner workings of more complex attention models, such as those used by *AlphaFold*. Moreover, we analyze the structure of the attention matrices to show their sparse nature in determining the structure of the protein family.

Another significant result found in [25] is that a factored model can be used to integrate signals from different protein families that can therefore share parameters. In particular, they showed that a set of shareable parameters can be learned from a protein family and then used with multiple other families for contact prediction, without loss in accuracy. Inspired by this, we introduce a multi-family learning scheme in which the shareable parameters are learned simultaneously on different protein families and used for contact prediction. This application exploits the factored nature of the model, highlighting the fundamental assumption that the signal from a protein family can be divided into two contributions: a family-specific signal and a universal, shareable signal that arises from structures and interactions common to all protein families.

In addition, we investigated the model's hyper-parameter space, finding that it is characterized by an iso-performance phase diagram defined by the overall number of parameters. More precisely, by defining H the number of heads and d the inner dimension of the factored attention model, we observed that on the H, d curve induced by fixing the number of parameters, the contact prediction's accuracy of the model is roughly constant (and improves upon increasing the total number of parameters, up to a certain threshold).

Finally, we introduced a generative version of the factored attention model by defining an autoregressive masking scheme inspired by the work in [21] where the same is applied to standard DCA. Sampling from the learned distribution produces generated MSAs that reproduce the same statistics as the natural ones. The effectiveness of our Attention-Based DCA architecture is assessed using different families of evolutionary-related proteins, whose alignments are sourced from the InterPro [27] (formerly Pfam [28]) database and structural data from the Protein Data Bank [29].

As shown in the Results Section, we argue that the attention mechanism and the Potts model are practically equivalent, in agreement with the recent theoretical evaluation using the Replica Method by Rende et al. [30]. Moreover, when modeling homology data, the transformer architecture gives possible advantages for analyzing the universal biochemical information which, if fully characterized and defined in terms of value matrices, could aid deep-learning machines such as *AlphaFold* itself.

Throughout the manuscript, we will use a Julia [31] in-house implementation of the factored self-attention mechanism that we refer to as *AttentionDCA*, to align this work to the terminology used in the field of DCA methods.

Methods

Factored attention

As in every Direct Coupling Analysis implementation, sequences in a protein family are represented by a Multiple Sequence Alignment and can be thought of as independent samples from a probability distribution that we model as a Gibbs-Boltzmann measure over a Hamiltonian function defined as a Potts model [7]:

$$P(\boldsymbol{a}) = \frac{1}{Z} e^{-H(\boldsymbol{a})} \quad , \tag{1}$$

$$H(\boldsymbol{a}) = -\sum_{i,j} J_{ij}(a_i, a_j) - \sum_{i=1}^{L} h_i(a_i) \quad , \tag{2}$$

where, $\mathbf{a} = (a_1, \ldots, a_L)$ is a sequence of L amino acids (a_i take value in an alphabet of 21 letters), J, and h are respectively the direct interaction tensor and the local field terms, while Z is a normalization constant, also known as partition function in the Statistical Physics jargon, given by:

$$Z = \sum_{\boldsymbol{a} \in \{1,\ldots,q\}^L} e^{-H(\boldsymbol{a})} \quad , \tag{3}$$

where $q = 21$ is the length of the amino acid dictionary, corresponding to the 20 natural amino acids plus a gap sign used during the alignment procedure in building the MSA. The inverse temperature β, which usually multiplies the energy term, is set equal to one, which is equivalent to implying its dependence directly inside terms J and h. More details on the standard implementation of DCA models are discussed in the Supplementary Material Sections A, B and C.

At this generic stage, tensor $J \in \mathbb{R}^{L,L,q,q}$ encodes both positional and amino acid information, while $h \in \mathbb{R}^{L,q}$ represents the local biases of each position in the sequence. To implement the factored attention mechanism discussed in [25], we discard the local field term, which is possible due to a gauge invariance of the parameters, and we write the interaction tensor mimicking the popular transformer [22] implementation of the attention mechanism for which:

$$J_{ij}(a_i, a_j) = \sum_{h=1}^{H} \texttt{softmax}(Q^h {K^h}^T)_{ij} V^h_{a_i,a_j} = \sum_{h=1}^{H} A^h_{ij} V^h_{a_i,a_j} \tag{4}$$

where, using the jargon from NLP, $Q^h \in \mathbb{R}^{L,d}$, $K^h \in \mathbb{R}^{L,d}$, $V^h \in \mathbb{R}^{q,q}$ are respectively the *query*, *key* and *value* matrices in one of H attention heads that build the interaction tensor J. The softmax is performed column-wise and the resulting matrix A^h_{ij}, $i, j \in 1, \ldots, L$ (such that for all $i \in 1, \ldots, L$, $\sum_{j=1}^{L} A^h_{ij} = 1$) represents the self-attention of each pair of residues. The attention matrix in this form is meant to highlight the co-evolution relationships between different positions in a sequence. Alternatively, phylogenetic relations between different sequences in the same MSA can be learned through a *column-attention* mechanism; more on this can be found in Rao et al. [32] and Sgarbossa at al. [33].

The structure of the factored attention mechanism shares similarities with the Hopfield-Potts model studied in [34–36], which also presents a low-rank decomposition of the Potts interaction tensor. However, a major difference between the Hopfield-Potts decomposition and that of Eq. 4 lies in the factorization between positional information contained inside matrices Q and K, depending on the specific protein family at hand, and the information contained inside matrix V which should capture the universal traits characterizing the interactions between the twenty natural amino acids. This turns out to be an interesting point for future development that plays an already crucial role in the multi-family version of the model discussed in the Multi- Family Learning subsection.

Another notable difference between the two models lies in their parameter inference techniques. The Hopfield-Potts model, which has the advantage of being analytically tractable, selects specific informative eigenmodes from the MSA's correlation function to infer its parameters. In contrast, the factored attention model employs a pseudo-likelihood approximation, which can be seen as a precursor to the more general Masked Language Modeling scheme, which is now widely used for training state-of-the-art Large Language Models [37]. The full probability distribution is given by a factorization into single-site distributions conditioned to all the other sites in the sequence. The pseudo-likelihood of the model over a Multiple Sequence Alignment (MSA) $\mathcal{D} = \{a_i^m\} \in \mathbb{R}^{L,M}$ with M sequences (depth of the MSA) each of L amino acids (length of the MSA) can be written as

$$\begin{aligned}\mathcal{L}(\mathbf{J}|\mathcal{D}) &= \frac{1}{M}\sum_{m=1}^{M}\sum_{i=1}^{L}\log P(\mathbf{a}^m|\mathbf{J}) = \\ &= \frac{1}{M}\sum_{m=1}^{M}\sum_{i=1}^{L}\left\{\sum_{j\neq i}J_{ij}(a_i^m,a_j^m) - \log\left[\sum_{a=1}^{q}\exp\left\{\sum_{j\neq i}J_{ij}(a,a_j^m)\right\}\right]\right\} \quad .\end{aligned} \tag{5}$$

Finally, following a general procedure and in agreement with what is used in PlmDCA [10], we add an L2-regularization of the interaction tensor after an analysis of different regularization schemes. This penalizes large values of the J_{ij} matrix to avoid overfitting the direct coupling score, as defined in Eq. 7. Thus, the total final likelihood is given by:

$$\tilde{\mathcal{L}}(\mathbf{J},\mathcal{D}) = \mathcal{L}(\mathbf{J},\mathcal{D}) + \lambda\sum_{i,j,a,b}J_{ij}^2(a,b) \quad . \tag{6}$$

The decomposition of the interaction tensor given by Eq. 4 has the side effect of making the total log-likelihood a non-convex-up function of its parameters, mainly due to the matrix products and the `softmax` function. Compared to models such as PlmDCA or ArDCA, this makes the maximization procedure much more challenging. To smooth out the complex landscape of peaks and troughs of the total likelihood during the maximization procedure, we implement an ADAM stochastic gradient ascent over mini-batches of fixed size [38].

Once the parameters are inferred, the standard way to obtain a contact prediction is by computing the *average-product-corrected* (cf. Supplementary Material Sec. C) Frobenius norm of the symmetrized interaction tensor which is interpreted as a score of the direct interaction between any position pair (i, j) in the chain [39]:

$$F_{ij} = \sqrt{\sum_{a_i,a_j} \tilde{J}_{ij}^2(a_i,a_j)} \quad , \tag{7}$$

where $\tilde{J}_{ij}(a_i,a_j) = \frac{1}{2}\Big(J_{ij}(a_i,a_j) + J_{ji}(a_j,a_i)\Big)$. The higher the Frobenius score, the more likely two pairs are to be an actual contact in the structure. This is indeed what is shown in [25]. However, as mentioned in the introduction, we implement an alternative scheme to compute a contact score in terms of the positional-dependent attention matrices. This attention score can be defined as

$$A_{ij} = \sum_{h=1}^{H} \mathrm{symm}\Big(A_{ij}^h\Big) = \sum_{h=1}^{H} \mathrm{symm}\Big(\texttt{softmax}(Q^h {K^h}^T)_{ij}\Big) \tag{8}$$

and it can be used in place of the original Frobenius score. This alternative attention score is motivated by its simplicity and direct connection to the attention mechanism, eliminating the need for complex or arbitrary definitions like the Frobenius score. As shown in the Results Section and thoroughly investigated analytically in Supplementary Material Sec. H, both methods exhibit comparable performance, with each offering advantages across different protein families.

The contact prediction accuracy of the model is evaluated by comparing the predicted contacts with a list of experimental contacts extracted from crystal structures from the Protein Data Bank database [29]. Following literature standards, two sites are considered in contact when the distance between the heavy atoms inside the amino acids is ≤ 8Å. Also, to avoid focusing on trivial contacts corresponding to sites close to each other in the amino acid chain, we consider only sites (i, j) such that $|i-j| \geq 6$. We define the Positive Predicted Value (PPV) as the percentage of true positive (TP) contacts among the predicted ones: $\mathrm{PPV}(n) = \mathrm{TP}(n)/n$. The curve given by this measure, a function of n, represents the global accuracy of the model in inferring the contact map of a protein family. Throughout the manuscript, we will use the consolidated notation PPV@n to indicate PPV(n).

Finally, due to the non-convex nature of the optimization problem and the related occurrence of many local stationary points, different training runs started from random initial set of parameters find slightly different solutions. For this reason, we perform multiple inference runs on the same protein family, producing a final Frobenius score by aggregating the results. Specifically, for each position pair (i, j), the merged score is taken as the maximum score across all iterations. Performing approximately 20 runs is sufficient to obtain a robust contact prediction. Furthermore, as discussed in Supplementary Material Sec. D, the variability in the contact score from single inference runs impacts only a subset of the predicted contacts, while a fixed core of contacts, whose number is proportional to the length of the protein family, remains consistently predicted. Figure S1 in the Supplementary Material shows the frequency of the core contacts through multiple runs for each protein family under study. In the Results Section we compare PPV curves from single runs and merged multiple runs of AttentionDCA.

Multi-family learning

A long-standing goal of protein design is that of being able to selectively pick features from different protein families and generate sequences at first in silico and then in vivo so that the new artificial proteins reproduce those specifically chosen traits and functions [40, 41]. The first step toward this feat is that of designing a DCA model that can learn simultaneously from different MSAs. Standard DCA is not prone to this possibility, since there is no obvious way to determine a set of parameters that can be shared among different families since both the interaction tensor and the local fields are family-dependent. However, in the Factored Attention implementation of DCA, the Value matrix $V \in \mathbb{R}^{q,q}$ does not depend on the specific family. In [25] it is shown that a set of Value matrices learned in a specific protein family can be *frozen* and used during the learning of Query and Key matrices for another protein family. This pre-training scheme is mentioned in Sec. H of the Supplementary Material as an argument in the analysis of the contribution of the Value matrices in the model. Expanding the work in [25], we introduce a novel multi-family model in which the Value matrices are shared and learned across different MSAs simultaneously. Given N_F protein families, a simple implementation of this is obtained by defining a multi-family likelihood $\mathcal{L}_{MF}$ given by the sum of the N_F single-family likelihoods coupled by the same set of Value matrices $\{V^h\}_{h=1,\dots,H}$:

$$\mathcal{L}_{MF} = \sum_{a=1}^{N_F} \mathcal{L}_a(\{Q_a^h, K_a^h, V^h\}, \mathcal{D}_a) \quad , \tag{9}$$

where each single-family likelihood is given by Eqs. 4 and 5. The set of parameters $\{Q_a^h, K_a^h\}$, one for each family, and $\{V^h\}$, shared across all families, can be used to extract the contact score of each MSA used during the learning. The $\{Q_a^h, K_a^h\}$ matrices must each have the same number of heads so that it is possible to share a common set of $\{V^h\}$ matrices among them. In the Results Section, we show the results of this PPV compared to those obtained from standard single-family learning. Pseudo-likelihood maximization is used to infer both the shared and family-specific parameters.

Generative model

The ability to efficiently generate realistic protein sequences is a key challenge in protein modeling, as it enables both functional predictions and protein design. The maximum pseudo-likelihood criterion used for inference in PlmDCA and AttentionDCA, however, neither provides a fast way to sample from the probability distribution nor ensures accurate sampling due to the factorization approximations. Because of this, to develop a generative version of the model we started from ArDCA [21], which is so far the state of the art for generative DCA architectures. In particular, it exploits a simple autoregressive model enforced by the exact decomposition

$$P(\mathbf{a}) = P(a_1) \prod_{i=2}^{L} P(a_i | a_1, \dots, a_{i-1}) \tag{10}$$

with

$$P(a_i|a_1,\dots,a_{i-1}) = \frac{1}{Z_i(a_{<i})} \exp\left\{\sum_{j<i} J_{ij}(a_i,a_j)\right\}, \tag{11}$$

$$Z_i(a_{<i}) = \sum_{a=1}^{q} \exp\left\{\sum_{j<i} J_{ij}(a,a_j)\right\} \quad . \tag{12}$$

This form allows for fast computation of the single-site partition function and a quick autoregressive sampling from L single-valued distributions.

To exploit the already existing libraries of ArDCA and the architecture for AttentionDCA, we apply a multiplicative mask to the interaction tensor to implement an autoregressive structure of the distribution, i.e. so that J_{ij} is a lower-triangular matrix with zeros above and on the diagonal:

$$J_{ij}(a_i,a_j) = \begin{cases} \sum_{h=1}^{H} \texttt{softmax}(Q^h {K^h}^T)_{ij} V^h_{a_i,a_j} & i > j \\ 0 & i \le j \end{cases} \quad . \tag{13}$$

As discussed in Trinquier et al. [21], the resulting interaction tensor cannot be interpreted as a matrix of direct couplings in the same way as in standard DCA, mainly because in the autoregressive implementation, the interactions of each position in the chain are conditioned only to partial sequences instead of being conditioned to all other amino acid positions. Therefore, the common technique to predict the residue-residue contacts is to extract coupling information directly from the epistatic score evaluated from the model. For amino acids b_i, b_j at positions (i, j), their epistatic score is defined as the difference between the effects of simultaneous mutations on both sites and the sum of the single site mutations when introduced in the wild-type $\mathbf{a} = (a_i, \dots, a_j)$ as in:

$$\begin{aligned} \Delta\Delta E_{ij}(b_i,b_j) = {} & \Delta E(a_i \to b_i, a_j \to b_j) + \\ & - \Delta E(a_i \to b_i) - \Delta E(a_j \to b_j) \quad , \end{aligned} \tag{14}$$

where the energy difference for a given mutation $a_i \to b_i$ is given by:

$$\begin{aligned} \Delta E(a_i \to b_i) &= E(a_1,...,b_i,...,a_L) - E(a_1,...,a_i,...,a_L) \\ &= -\log \frac{P(a_1,...,b_i,...,a_L)}{P(a_1,...,a_i,...,a_L)} \quad . \end{aligned} \tag{15}$$

The $\Delta\Delta E_{ij}(b_i,b_j)$ replaces the interaction tensor $J_{ij}(a_i,a_j)$ in the Frobenius norm, cf. Equation 7, used to build the Epistatic Score ES_{ij} so that a contact map can be produced and compared with the true structure of the protein. It is worth mentioning that the choice of the wild-type sequence in the definition of the Epistatic Score (ES) is practically immaterial. In Supplementary Material Sec. E, we discuss how different choices of the wild-type reference sequence, including randomly generated ones, produce equivalent results. This behavior, first noted in the context of ArDCA, highlights that the relevant evolutionary information is fully encoded within the interaction tensor. As such, the ES efficiently captures this information, regardless of the reference sequence used. Figure S2 in the Supplementary Material shows that the PPV@L for each family remains unaffected by the choice of the wild-type sequence, underscoring the robustness of

this approach. Finally, an alternative to using the Epistatic Score to produce a contact score from the generative model is that of sampling an artificial MSA to be fed directly to a non-generative version, either PlmDCA or AttentionDCA, to define a meaningful Frobenius Score. Both alternatives will be discussed in the next section.

Results

The model, in its standard and autoregressive implementations, has been tested against nine protein families whose structural data was taken from the Protein Data Bank [29]. Each family is represented by an MSA whose length L (number of amino acids per sequence) and depth M (number of sequences per MSA). Table 1 summarizes some of the main information regarding each family. These families were chosen to test the model against a variable dataset in terms of length, depth and effective depth M_{eff}, i.e. the number of unique non-redundant sequences in the MSA as discussed in the Supplementary Material Sec. B. As we show in the following, the effective depth turns out to be crucial in determining the accuracy of a model and in particular the contact prediction accuracy of the autoregressive generative version of the model, cf. subsection Autoregressive Generative Version.

The model is evaluated in its current implementations for contact prediction and in silico sequence generation. Future research will explore the performance and limitations of AttentionDCA and the factored attention mechanism in other key applications of DCA, such as mutational landscape inference and protein-protein interaction modeling.

Standard version

Parameter reduction and hyper-parameters

Concerning the hyper-parameters of the learning process, we use a learning rate $\eta = 0.005$ and a minibatch size $n_b = 1000$, while the number of epochs varies depending on the depth of each MSA, cf. Table S1 in the Supplementary Material for details about the choice of parameters for each family. A more insightful analysis regards the hyper-parameters of the model itself, i.e. the number of heads H used to define the interaction tensor and the inner dimension d of the rectangular matrices Q and K. In standard DCA models, the size of the interaction tensor scales quadratically with the length of the protein family. However, empirically, the number of contacts in a protein is proportional to its length L (and not to L^2) [42], therefore we can expect the interaction tensor to be effectively sparse. In AttentionDCA, due to the low-rank decomposition of the interaction tensor, the scaling of the parameters is linear with the size of the family L: $N_{\text{AttentionDCA}} = 2HLd + Hq^2$; while for PlmDCA $N_{\text{PlmDCA}} = L(L-1)q^2/2 + Lq$. The parameter compression ratio relative to PlmDCA can be computed by fixing the value of d and H:

$$c_r := \frac{N_{\text{AttentionDCA}}}{N_{\text{PlmDCA}}} = \frac{2HdL + Hq^2}{L(L-1)q^2/2 + Lq} \quad . \tag{16}$$

While tuning the model's hyper-parameters, we observe that, at a fixed compression ratio, the precision of the contact prediction is constant regardless of the choice

of H and d. Therefore the number of parameters determines the quality of the model. Although a similar analysis regarding the dependence of the contact prediction's accuracy on the number of heads of the model is present in [25], our findings show that it isn't the number of heads that influences the accuracy of the model, but its global number of parameters. However, increasing the compression ratio, i.e., lowering the distance between PlmDCA and AttentionDCA in terms of the number of parameters, does not improve the contact prediction's accuracy indefinitely. Indeed, having the same number of parameters of PlmDCA results in sensibly worse results. This may be because the likelihood in this architecture does not have an absolute maximum due to the function's non-convex nature. Increasing the parameters exacerbates this condition, making inference progressively more difficult. After some trials, we conclude that the optimal parameter compression lies between 5% and 20%. Focusing on families PF00014, PF00076, and PF00763 respectively, Fig. 1 shows the PPV evaluated at L, $3L/2$, and $2L$ for different values of the head number H and the inner dimension d so that the parameter compression is fixed at values $c_r = \{0.05, 0.15, 0.25, 0.35\}$. It can be seen how a plateau is reached for $H > 10$, effectively highlighting the fact that performances remain constant at fixed compression, regardless of the hyper-parameters H and d. Another interesting feature is that for $H \lesssim 10$, the value of the PPV is significantly lower than otherwise. This is due to a non-reliable inference of the parameters, as it can be seen that soon after the beginning of the gradient ascent, it hits a barrier and stops. The results shown in Fig. 1 are averaged across multiple realizations, even though for most points the error bar is too small to be noticeable.

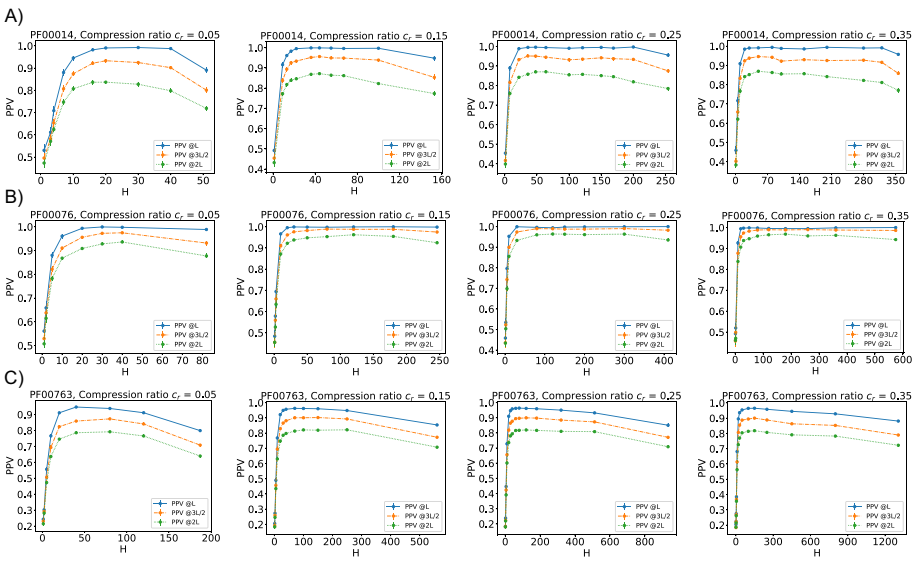

Fig. 1 PPV@L, 3 L/2 and 2 L (blue, orange and green curve respectively) for different values of the compression ratio r for families PF00014, PF00076 and PF00763 respectively in panels (**A**), (**B**) and (**C**). In each plot the compression ratio is held constant while the Positive Predicted Value is calculated across different inference runs with varying numbers of heads H. Since the compression ratio is fixed, for each value of H there is a corresponding value of d given by inverting Eq. 16 at fixed c_r

Standard and multi-family contact prediction

The standard way to perform a contact prediction for DCA methods is to compute the Frobenius norm of the inferred and symmetrized interaction tensor after a transformation which brings the parameters in zero-sum gauge, the gauge that minimizes the Frobenius norm, and a consecutive *average product correction* [10] (discussed in the Supplementary Material Sec. C). Sorting the Frobenius score (FS) results in a list of contacts that can be compared to the actual known structure of the protein family to produce a PPV curve. Alternatively, the factored attention form of the interaction tensor suggests another way of determining the contacts of a protein family. Averaging the positional-dependent symmetrized attention matrix through each head results in a matrix of the form given by Eq. 8, the attention score (AS). After the *average product correction*, the contact prediction is given by the list of contacts (i, j) sorted by highest A_{ij}. As already mentioned, the non-convexity of the problem and the roughness of the energy landscape produce a noise in the final inference which can be smoothed out by averaging through different realizations. In particular in the case of the FS contact prediction we compute the regular Frobenius score for each realization and define the merged score for each position pair (i, j) as the maximum score through all realizations: $F_{ij} = \max\{F_{ij}^1, F_{ij}^2, \ldots, F_{ij}^m\}$. This approach yields a final, merged score that reflects the most likely outcome. In the case of the AS contact prediction, it is sufficient to average the resulting attention matrices of each realization. Following an analysis with an increasing number of realizations m, we identified an accuracy threshold at $m = 20$, which was subsequently used for the simulations.

Figure 2 shows the comparison between the PPV curves extracted from different implementations of AttentionDCA and those from PlmDCA (black curve), which serves as a benchmark model, applied to families PF00014, PF00072 and PF00763. Gray curves represent an ideal perfect model which predicts all and only positive contacts, while the red and blue curves are respectively the FS PPV from the merged- and single-realization of AttentionDCA. These results have been obtained by setting the compression ratio between PlmDCA and AttentionDCA to 20%, and for each family, apart from fluctuations in the single-realization curves, the results are in very strong agreement with PlmDCA, despite the significant parameter reduction between the two models.

In the multiple-family learning approach, we observed that the results achieved for contact prediction were consistent with those obtained in single-family learning.

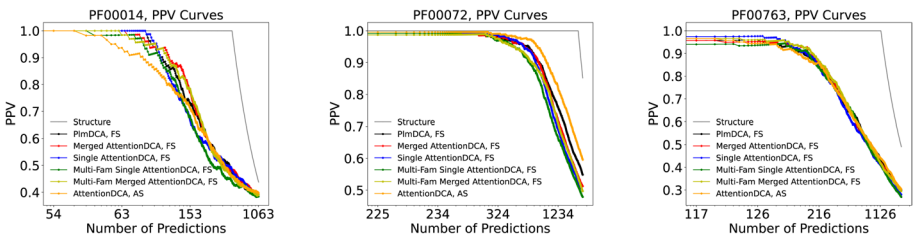

Fig. 2 Positive Predicted Value Curves for families PF00014, PF00072, PF00763 respectively. Each plot compares PPVs computed through the Frobenius Score (FS) and the Attention Score (AS), in particular: PlmDCA with FS (black curve); merged- and single-realization of AttentionDCA with FS (red and blue curves respectively); Multi-Family AttentionDCA with merged- and single-realization (green and yellow curves respectively); AttentionDCA with AS (orange curve). The gray curve represents the PPV of an ideal model that predicts all and only positive contacts, hence it depicts the prediction of the full structure of the family

Complemented by the analysis on *frozen* Value matrices in [25] and replicated in Supplementary Material Sec. H., this indicates that the shared parameter set effectively captures the essential bio-chemical information across different protein families without compromising prediction accuracy. By training on multiple families with shared Value matrices, the model leverages common interaction patterns while retaining family-specific features through separate Query and Key matrices. This setup not only maintains accuracy comparable to single-family models but also highlights the model's flexibility in integrating diverse protein family data, thus offering an efficient alternative for multi-family protein analysis. Dark and light green curves in Fig. 2 represent the PPVs evaluated from the multi-family leaning discussed in subsection Multi-Family Learning. In particular, the learning has been performed in parallel among all families using a fixed number of heads $H = 128$ and varying the inner dimension d to match the single-family learnings at 20% compression ratio. The shown curves are obtained through the Frobenius score computed on the J_{ij} tensor which is given by Eq. 4 and parameters $\{Q_F^h, K_F^h V_{\text{shared}}^h\}_{h=1,\dots,H}$, for $F = \{\text{PF00014}, \text{PF00072}, \text{PF00763}\}$. Figure S7 in the Supplementary Material depicts the same single- and multi-family analysis for the remaining protein families under study.

Attention heads

In the field of Natural Language Processing [43], the self-attention mechanism is used to learn custom functional relationships between the elements of a dataset to produce an encoded description of the input itself [44]. In the context of protein structure prediction, the attention mechanism used in *AlphaFold* untangles the physical and evolutionary constraints between phylogenetically related sequences [45, 46]. Since the factored attention model is effectively a single-layer self-attention architecture, we expect the attention heads inside the J_{ij} tensor to directly capture some level of structural information that could be compared to the actual contact map obtained by the average-product-corrected Frobenius norm of the interaction tensor. Ideally, one would expect each head to focus its attention on a specific spatial structure in the same way as a transformer head would specialize in different semantic functions when used in the context of NLP [47]. However, it is extremely difficult to determine if this is the case for protein attention due to the lack of an obvious interpretation key between the several possible structures arising in a protein and the variable number of heads. Also, given the form of Eq. 4, a single head generally produces an asymmetric attention matrix, hence it is hard to understand which are the structures emerging from each one of them. A solution to both these problems is to simply average out all heads into a single attention matrix as in Eq. 8.

The average attention matrix can be used to extract structural information by producing a list of contacts as discussed in subsection Standard and Multi-Family Contact Prediction and shown by the orange curves in Fig. 2. However, the most direct way in which the average attention matrix can be interpreted is by comparing it to the actual contact map of the protein family. Figure 3 compares the contact map built from a significant fraction of predicted contacts and the heat map of the attention matrix for families PF00014, PF00072, and PF00763. In the contact maps, blue and red dots represent the positive and negative predictions, while gray dots show the actual structure of the family. Conversely, the heat maps are displayed on a grayscale, while the actual structure is given

by the red dots underneath. It can be seen that the maxima of the attention matrix, i.e. the darker dots, lie within the structure of the protein, proving that the attention model is focused on the structural contacts. As shown in Figures S8, S9 and S10 of the Supplementary Material, a similar conclusion can be reached from other protein families.

Even though, as already mentioned, the specialization of individual attention heads is challenging to interpret, an important observable feature is the sparsification of structural information within each head. Supplementary Material Sec. G provides a quantitative analysis of this sparsity. Specifically, Figures S3 and S4 demonstrate that utilizing only a small fraction of the information from each attention head is sufficient to achieve contact prediction accuracy comparable to that of the full attention matrix. Figure 3C) shows the heatmap of the sparse attention matrix constructed by maximizing over the k highest elements from each head, using $k = L$.

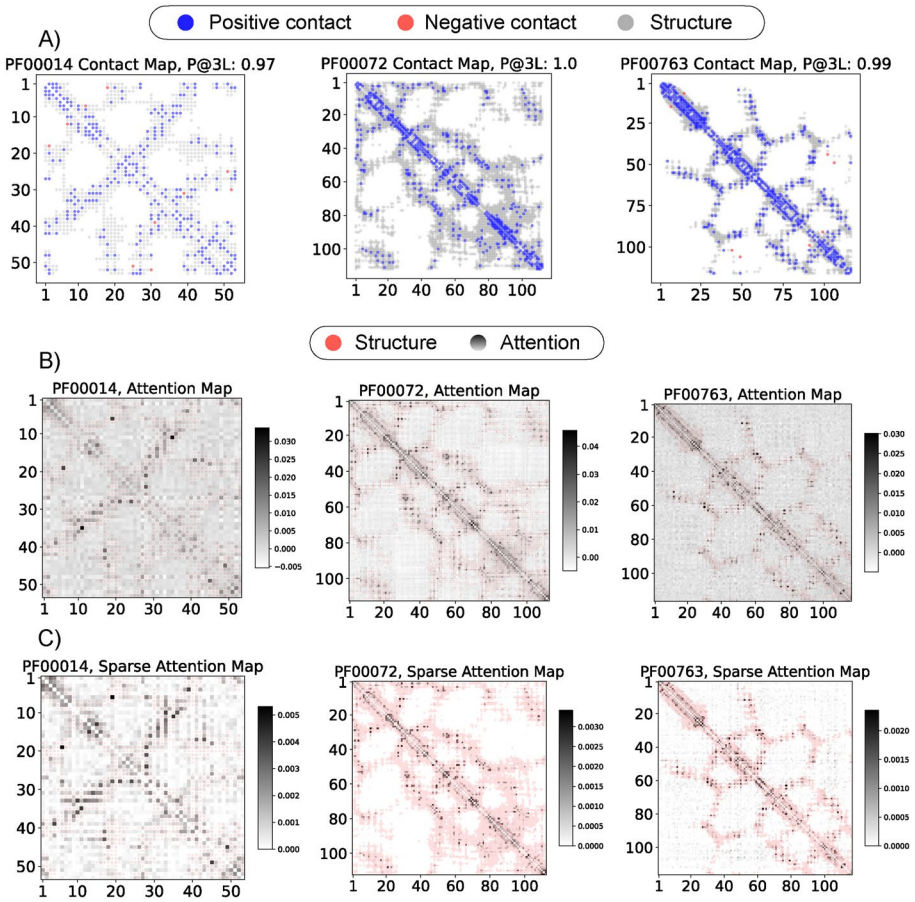

Fig. 3 Various contact maps for families PF00014, PF00072, PF00763. **A** Contact map built from the Frobenius Score applied to the interaction tensor inferred by AttentionDCA. Blue and red dots represent positive and negative predictions respectively, while gray dots reproduce the structure of the family. **B** Grayscale heat map of the averaged attention matrix built using Eq. 8, from the $\{Q^h, K^h\}_{h=1,\dots,H}$ matrices by AttentionDCA. **C**) Grayscale heat map for the sparse attention matrix constructed by maximizing over the k highest elements from each head, using $k = L$, cf. Supplementary Material Sec. G. Both in **B**) and **C**) structural information is depicted in red

Autoregressive generative version

Even though the autoregressive version of our architecture can be used for contact prediction, its main target is to efficiently sample the distribution to generate large MSAs matching the summary statistics of the natural MSAs used for inference.

As described in the Generative Model Section, the form of the interaction term $J_{ij}(a, b)$ of the autoregressive version does not allow for a meaningful mapping to direct interactions among amino acids. Therefore, the epistatic score measure defined in Eq. 14 must be used to define a list of predicted contacts and the corresponding PPV curve. Figure 4A) shows different PPV curves computed on families PF00014, PF00072 and PF000763. In particular, it compares the performance of PlmDCA (black curve) and ArDCA (red curve), both acting as benchmarks, and those of the autoregressive AttentionDCA in the single- and merged-realization implementations (blue and green curves respectively). Along with the epistatic measure, another way to perform a contact prediction is to produce a generated MSA, sampled using AttentionDCA, and feed it to PlmDCA or AttentionDCA itself (in its non-generative version). This last possibility is represented by the yellow curve and it is among the best options for contact prediction within the autoregressive versions, either AttentionDCA or ArDCA. This can also be thought of as a way to test the accuracy of the generative capabilities of the autoregressive model, which exploits the exact decomposition in Eq. 10 to sample from L single-valued probability distributions. Moreover, the generativity can be assessed by comparing the summary statistics of natural and generated MSAs. The most informative summary statistics are given by the connected two-site correlations of amino acids in different positions of an MSA, defined in Supplementary Material Sec. A. Their comparison is shown in Fig. 4B), along with their Pearson's correlation coefficient, proving the accuracy of the model. The analysis on contact prediction and generativity of the autoregressive model is expanded on the remaining protein families in Figures S11 and S12 in the Supplementary Material. Finally, a PCA analysis can be performed to show that the principal components of the generated data are comparable to those of the natural data. Figure 5 shows the comparison between the two principal components of the natural data and the artificial data generated by ArDCA, used as a benchmark model, and by AttentionDCA on families PF00014, PF00072 and PF13354 (cf. Supplementary Material Figure S13 for the same analysis on other families).

For many families among the ones described in Table 1, the results of the contact prediction from the epistatic score lie significantly below those from the Frobenius score of the standard version of AttentionDCA or PlmDCA. The reasons for this can be found in the dependency of the autoregressive version on the M_{eff} of the family. If one were to reduce the effective depth of an MSA, after a certain threshold, the DCA inference would suffer a lack of observations and the contact prediction would get less and less reliable. Different DCA models and different inference techniques have variable M_{eff} thresholds.

Figure 6 shows how the accuracy of different DCA methods decays as a function of the effective depth for families of various lengths. While PlmDCA exhibits the best resilience when the effective depth is lowered, having an adequate accuracy even below

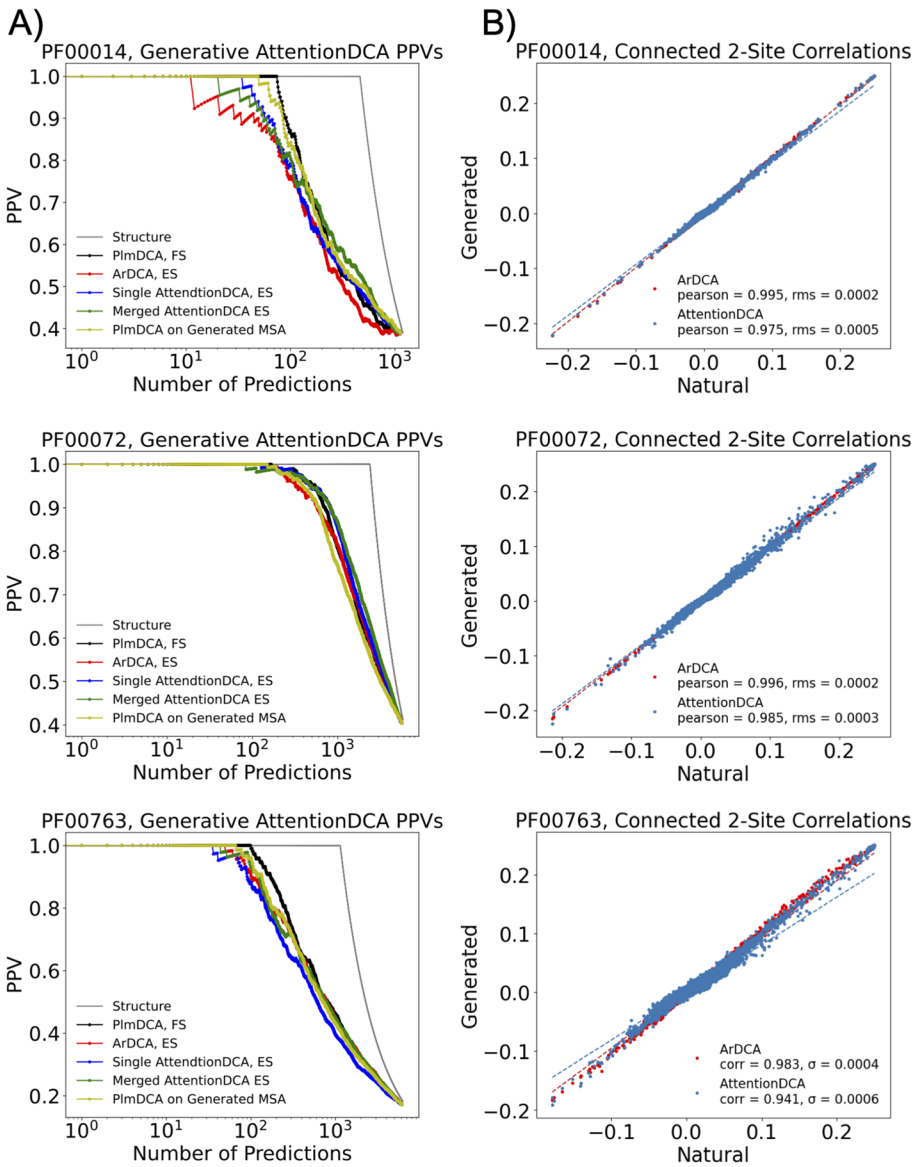

Fig. 4 Results from Autoregressive AttentionDCA applied to families PF00014, PF00072 and PF00763. ES stands for Epistatic Score. For each family, the panels show the PPV curves and the connected two-site correlations. **A** The plots on the left compare the PPV curves from PlmDCA (FS, black curve), ArDCA (ES, red curve) with the single- and merged-realizations of AttentionDCA with ES (blue and green curves respectively) and the PPV from the FS of PlmDCA applied to a Generated MSA sampled from Autoregressive AttentionDCA. The gray curve represents the PPV of an ideal model that predicts all and only positive contacts, hence it depicts the prediction of the full structure of the family. **B** The plots on the right show a comparison between the two-site connected correlations of the natural and artificial MSAs generated by Autoregressive AttentionDCA and ArDCA (blue and red dots respectively)

$M_{\text{eff}} < 1000$, the autoregressive version of AttentionDCA degrades for $M_{\text{eff}} \leq 2000$. Because of this, performances on shallow families such as PF00595, PF13354 or PF00035 are intrinsically worse than those from PlmDCA and the standard version of AttentionDCA.

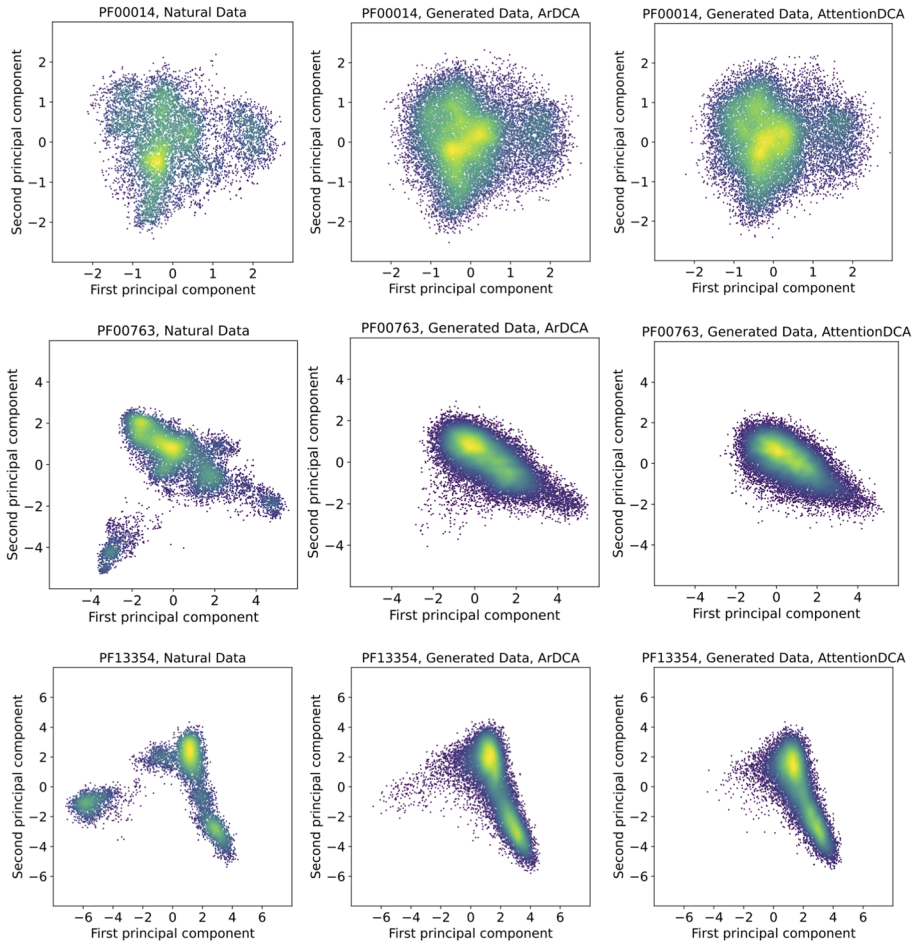

Fig. 5 Principal Component Analysis on protein families PF00014, PF00763 and PF13354. Comparison between natural data and artificial data generated by using ArDCA and AttentionDCA

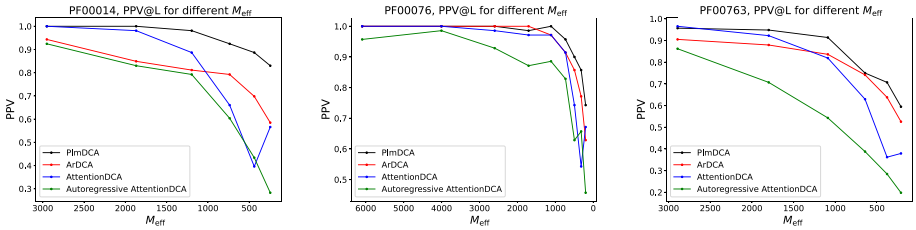

Fig. 6 PPV@L as the M_{eff} of the family varies, computed using PlmDCA (FS), ArDCA (ES), AttentionDCA (FS) and Autoregressive AttentionDCA (ES). Families PF00014, PF00076, and PF00763 are shown

Discussion

During the past decade, since the development of DCA methods, it became evident that parameter reduction was essential to avoid overfitting the limited amount of information available in MSAs. Early approaches addressed this issue by adopting low-rank decomposition techniques [36], enabling models to represent the high-dimensional interaction tensors more compactly while retaining the essential coevolutionary signals. Similarly, sparse models [48, 49] introduced regularization

Table 1 Summary of the protein families used throughout the analysis and main results

Name	Description	L	M	M_{eff}	N_{PlmDCA}	$N_{AttentionDCA}$	ρ_{ArDCA}	$\rho_{AttentionDCA}$	PPV@L PlmDCA	PPV@L AttentionDCA
PF00014	Kunitz Domain	53	8871	2950	608,811	124,288	0.995	0.975	1.0	1.0
PF00035	DSRM	67	18462	2176	976,458	193,664	0.986	0.963	0.880	0.895
PF00072	RR	112	574565	59290	2,743,608	543,872	0.996	0.985	1.0	1.0
PF00076	RRM	70	79366	6093	1,066,485	217,728	0.997	0.986	1.0	0.985
PF00169	PH Domain	105	55196	4115	2,410,065	486,528	0.985	0.958	0.990	0.990
PF00595	PDZ Domain	82	15299	1489	1,466,283	287,360	0.973	0.846	0.939	0.878
PF00677	Lumazine BD	87	16192	2853	1,651,608	323,712	0.994	0.982	0.908	0.896
PF00763	Catalytic Domain	116	10886	2892	2,943,906	590,976	0.983	0.941	0.956	0.956
PF13354	Beta-Lactamase	202	7515	1559	8,956,983	1,814,656	0.942	0.910	0.965	0.935

Name, Description, Length, Depth, Effective Depth of the Protein Family. Number of parameters for PlmDCA and AttentionDCA. Two-site connected correlation Pearson coefficient ρ for ArDCA and AttentionDCA. FS Positive Predicted Values at L for PlmDCA and AttentionDCA

schemes to focus on the most informative residue-residue interactions, further mitigating overfitting and improving the robustness of contact predictions. More recently, collective efforts to enhance DCA have increasingly incorporated machine learning techniques and attention mechanisms. These advancements range from shallower implementations, such as the one discussed in [25], to fully deep models leveraging transformer architectures [32, 50–52], ultimately culminating in breakthroughs such as AlphaFold [2].

In this study, we present a comprehensive analysis of the single-layer factored self-attention mechanism, offering a thorough comparison with the standard pseudo-likelihood approximation used in PlmDCA. Our exploration spanned a diverse dataset, encompassing nine protein families meticulously curated from the Protein Data Bank. The primary feature of a factored attention layer compared to a Potts model is the low-rank decomposition that ensures a significant parameter reduction without sacrificing structure prediction accuracy. In particular, the attention mechanism learns structural determinants directly at the level of the attention matrices which can be directly used as a contact score, as shown in the Standard and Multi-Family Contact Prediction subsection. An additional strength of the factored attention mechanism lies in its ability to differentiate between positional and amino acid epistatic signals in an MSA. As described in the Multi-Family Learning subsection, this leads to the possibility of integrating universal amino acid interaction information from different protein families by sharing common sets of parameters in what represents the first example of a multi-family DCA method.

While the parameter reduction and the multi-family modeling of the current implementation do not enhance the contact prediction or the computational complexity of the model, they offer a theoretical improvement and a novel flexible framework that can be adapted and tailored for specific biological problems. Applications may include analyzing interactions between structural or functional protein subdomains, mapping interactions between different proteins, reconstructing phylogenetic relationships, and exploring mutational landscapes.

Beyond its primary application in contact prediction, the factored attention layer seamlessly transitions into a fully generative architecture by implementing an autoregressive masking scheme, cf. subsection Generative Model. The heterogeneities observed in the principal component analysis of the generated datasets and the results in both contact prediction and sequence generation tests prove the versatility of our autoregressive factored attention model across various tasks, as shown in Figs. 4 and 5. Although the generative performance of AttentionDCA does not surpass that of similar architectures such as ArDCA [21], this model provides valuable insights into the generative properties of the factored attention mechanism and lays the groundwork for its potential applications in future studies.

The theoretical and experimental findings from [25] and [30], along with our in-depth analysis of the potentiality of the factored attention mechanism for protein contact prediction and sequence generation, constitute a step forward in bridging the gap between the remarkable achievements of attention mechanisms in computational biology and the depth of our understanding about their inner workings.

Among the many challenges for future developments, it will be crucial to understand the biological information that can be stored inside Value matrices and how this varies when learned from single or multiple protein families. Likewise, it will be interesting to determine whether single attention-heads can focus on specific structures or if it is always needed to integrate the information from all heads to produce a meaningful representation of a protein.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12859-025-06062-y.

Supplementary file 1.

Acknowledgements
The authors acknowledge illuminating discussions with Sergey Ovchinnikov, Martin Weigt, and Francesco Zamponi.

Author contributions
FC implemented the AttentionDCA model and conducted the computational experiments. AP supervised the study, provided guidance on the theoretical framework, and contributed to data interpretation. Both authors contributed equally to the design of the study, the analysis of results, and the writing of the manuscript. All authors read and approved the final manuscript.

Funding
The Authors acknowledge financial support from the project "Explainable Models for Protein Design", funded by the MIUR Progetti di Ricerca di Rilevante Interesse Nazionale (PRIN) Bando 2022 - grant 2022TE5B7X. We also acknowledge "Centro Nazionale di Ricerca in High-Performance Computing, Big Data, and Quantum Computing" (ICSC), and "FAIR - Future Artificial Intelligence Research", received funding from the European Union Next-GenerationEU (Piano Nazionale di Ripresa e Resilienza-Missione 4 Componente 2, Investimento 1.3-D.D. 1555 11/10/2022, PE00000013). This manuscript reflects only the authors' views and opinions, neither the European Union nor the European Commission can be considered responsible for them.

Availability of data and materials
Our in-house Julia implementation of the factored attention mechanism, also known as ***AttentionDCA***, and the MSAs and protein structure data used during experiments are available at github.com/pagnani/AttentionDCA.jl and github.com/francescocaredda/DataAttentionDCA respectively.

Declarations

Ethics approval and consent to participate
Not applicable

Consent for publication
Not applicable

Competing interests
No Conflict of interest is declared.

Received: 5 November 2024 Accepted: 22 January 2025
Published online: 06 February 2025

References

1. Dill KA, MacCallum JL. The protein-folding problem, 50 years on. Science. 2012;338(6110):1042–6. https://doi.org/10.1126/science.1219021.
2. 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 SAA, Ballard AJ, 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 AW, Kavukcuoglu K, Kohli P, Hassabis D. Highly accurate protein structure prediction with AlphaFold. Nature. 2021;596(7873):583–9. https://doi.org/10.1038/s41586-021-03819-2.
3. de Juan D, Pazos F, Valencia A. Emerging methods in protein co-evolution. Nat Rev Genet. 2013;14(4):249–61. https://doi.org/10.1038/nrg3414.
4. Cocco S, Feinauer C, Figliuzzi M, Monasson R, Weigt M. Inverse statistical physics of protein sequences: a key issues review. Rep Prog Phys. 2018;81(3): 032601. https://doi.org/10.1088/1361-6633/aa9965.
5. Durbin R, Eddy SR, Krogh A, Mitchison G. Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids. Cambridge University Press, Cambridge 1998. https://doi.org/10.1017/CBO9780511790492
6. Weigt M, White RA, Szurmant H, Hoch JA, Hwa T. Identification of direct residue contacts in protein-protein interaction by message passing. Proc Natl Acad Sci U S Am. 2009;106(1):67–72. https://doi.org/10.1073/pnas.0805923106.

7. Morcos F, Pagnani A, Lunt B, Bertolino A, Marks DS, Sander C, Zecchina R, Onuchic JN, Hwa T, Weigt M. Direct-coupling analysis of residue coevolution captures native contacts across many protein families. Proce Nat Acad Sci. 2011;108(49):1293–301. https://doi.org/10.1073/pnas.1111471108.
8. Marks DS, Colwell LJ, Sheridan R, Hopf TA, Pagnani A, Zecchina R, Sander C. Protein 3D Structure Computed from Evolutionary Sequence Variation. PLOS ONE. 2011;6(12):28766. https://doi.org/10.1371/journal.pone.0028766.
9. Baldassi C, Zamparo M, Feinauer C, Procaccini A, Zecchina R, Weigt M, Pagnani A. Fast and accurate multivariate Gaussian modeling of protein families: predicting residue contacts and protein-interaction partners. PLOS ONE. 2014;9(3):92721. https://doi.org/10.1371/journal.pone.0092721.
10. Ekeberg M, Lövkvist C, Lan Y, Weigt M, Aurell E. Improved contact prediction in proteins: using pseudolikelihoods to infer Potts models. Phys Rev E. 2013;87(1): 012707. https://doi.org/10.1103/PhysRevE.87.012707.
11. Gueudré T, Baldassi C, Zamparo M, Weigt M, Pagnani A. Simultaneous identification of specifically interacting paralogs and interprotein contacts by direct coupling analysis. Proc Nat Acad Sci. 2016;113(43):12186–91. https://doi.org/10.1073/pnas.1607570113.
12. Bitbol A-F, Dwyer RS, Colwell LJ, Wingreen NS. Inferring interaction partners from protein sequences. Proc Nat Acad Sci. 2016;113(43):12180–5. https://doi.org/10.1073/pnas.1606762113.
13. Hopf TA, Colwell LJ, Sheridan R, Rost B, Sander C, Marks DS. Three-dimensional structures of membrane proteins from genomic sequencing. Cell. 2012;149(7):1607–21. https://doi.org/10.1016/j.cell.2012.04.012.
14. Morcos F, Jana B, Hwa T, Onuchic JN. Coevolutionary signals across protein lineages help capture multiple protein conformations. Proc Nat Acad Sci. 2013;110(51):20533–8. https://doi.org/10.1073/pnas.1315625110.
15. Fernandez-de-Cossio-Diaz J, Uguzzoni G, Pagnani A. Unsupervised inference of protein fitness landscape from deep mutational scan. Mol Biol Evol. 2021;38(1):318–28. https://doi.org/10.1093/molbev/msaa204.
16. De Leonardis M, Fernandez-de-Cossio-Diaz J, Uguzzoni G, Pagnani A. Unsupervised modeling of mutational landscapes of adeno-associated viruses viability. BMC Bioinf. 2024;25(1):229. https://doi.org/10.1186/s12859-024-05823-5.
17. Sesta L, Pagnani A, Fernandez-de-Cossio-Diaz J, Uguzzoni G. Inference of annealed protein fitness landscapes with AnnealDCA. PLOS Comput Biol. 2024;20(2):1011812. https://doi.org/10.1371/journal.pcbi.1011812.
18. Hopf TA, Ingraham JB, Poelwijk FJ, Schärfe CPI, Springer M, Sander C, Marks DS. Mutation effects predicted from sequence co-variation. Nat Biotechnol. 2017;35(2):128–35. https://doi.org/10.1038/nbt.3769.
19. Ackley DH, Hinton GE, Sejnowski TJ. A learning algorithm for boltzmann machines. Cognit Sci. 1985;9(1):147–69. https://doi.org/10.1016/S0364-0213(85)80012-4.
20. Figliuzzi M, Barrat-Charlaix P, Weigt M. How pairwise coevolutionary models capture the collective residue variability in proteins? Mol Biol Evol. 2018;35(4):1018–27. https://doi.org/10.1093/molbev/msy007.
21. Trinquier J, Uguzzoni G, Pagnani A, Zamponi F, Weigt M. Efficient generative modeling of protein sequences using simple autoregressive models. Nat Commun. 2021;12(1):5800. https://doi.org/10.1038/s41467-021-25756-4.
22. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser L, Polosukhin I. Attention Is All You Need. arXiv 2017. https://doi.org/10.48550/arXiv.1706.03762
23. Nambiar A, Liu S, Hopkins M, Heflin M, Maslov S, Ritz A. Transforming the Language of Life: Transformer Neural Networks for Protein Prediction Tasks. bioRxiv 2020. https://doi.org/10.1101/2020.06.15.153643
24. Wolf T, Debut L, Sanh V, Chaumond J, Delangue C, Moi A, Cistac P, Rault T, Louf R, Funtowicz M, Davison J, Shleifer S, von Platen P, Ma C, Jernite Y, Plu J, Xu C, Le Scao T, Gugger S, Drame M, Lhoest Q, Rush A. Transformers: State-of-the-Art Natural Language Processing. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, pp. 38–45. Association for Computational Linguistics, Online 2020. https://doi.org/10.18653/v1/2020.emnlp-demos.6
25. Bhattacharya N, Thomas N, Rao R, Dauparas J, Koo PK, Baker D, Song YS, Ovchinnikov S. Single Layers of Attention Suffice to Predict Protein Contacts. bioRxiv 2020. https://doi.org/10.1101/2020.12.21.423882
26. Wu FY. The Potts model. Rev Mod Phys. 1982;54(1):235–68. https://doi.org/10.1103/RevModPhys.54.235.
27. ...Paysan-Lafosse T, Blum M, Chuguransky S, Grego T, Pinto BL, Salazar GA, Bileschi ML, Bork P, Bridge A, Colwell L, Gough J, Haft DH, Letunić I, Marchler-Bauer A, Mi H, Natale DA, Orengo CA, Pandurangan AP, Rivoire C, Sigrist CJA, Sillitoe I, Thanki N, Thomas PD, Tosatto SCE, Wu CH, Bateman A. InterPro in 2022. Nucleic Acids Res. 2023;51(D1):418–27. https://doi.org/10.1093/nar/gkac993.
28. Mistry J, Chuguransky S, Williams L, Qureshi M, Salazar GA, Sonnhammer ELL, Tosatto SCE, Paladin L, Raj S, Richardson LJ, Finn RD, Bateman A. Pfam: The protein families database in 2021. Nucleic Acids Res. 2021;49(D1):412–9. https://doi.org/10.1093/nar/gkaa913.
29. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE. The protein data bank. Nucleic Acids Res. 2000;28(1):235–42. https://doi.org/10.1093/nar/28.1.235.
30. Rende R, Gerace F, Laio A, Goldt S. What Does Self-Attention Learn from Masked Language Modelling? arXiv 2024. https://doi.org/10.48550/arXiv.2304.07235
31. Bezanson J, Edelman A, Karpinski S, Shah VB. Julia: A Fresh Approach to Numerical Computing. arXiv 2015. https://doi.org/10.48550/arXiv.1411.1607
32. Rao R, Liu J, Verkuil R, Meier J, Canny JF, Abbeel P, Sercu T, Rives A. MSA Transformer bioRxiv. 2021. https://doi.org/10.1101/2021.02.12.430858.
33. Sgarbossa D, Lupo U, Bitbol A-F. Generative power of a protein language model trained on multiple sequence alignments. eLife. 2023;12:79854. https://doi.org/10.7554/eLife.79854.
34. Cocco S, Monasson R, Sessak V. High-dimensional inference with the generalized Hopfield model: principal component analysis and corrections. Phys Rev E. 2011;83(5): 051123. https://doi.org/10.1103/PhysRevE.83.051123.
35. Cocco S, Monasson R, Weigt M. From principal component to direct coupling analysis of coevolution in proteins: low-eigenvalue modes are needed for structure prediction. PLOS Comput Biol. 2013;9(8):1003176. https://doi.org/10.1371/journal.pcbi.1003176.
36. Shimagaki K, Weigt M. Selection of sequence motifs and generative Hopfield-Potts models for protein families. Phys Rev E. 2019;100(3): 032128. https://doi.org/10.1103/PhysRevE.100.032128.

37. Devlin J, Chang M.-W, Lee K, Toutanova K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In: Burstein, J., Doran, C., Solorio, T. (eds.) 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), pp. 4171–4186. Association for Computational Linguistics, Minneapolis, Minnesota 2019. https://doi.org/10.18653/v1/N19-1423
38. Kingma DP, Ba J. Adam: a method for stochastic optimization. arXiv 2017
39. Dunn SD, Wahl LM, Gloor GB. Mutual information without the influence of phylogeny or entropy dramatically improves residue contact prediction. Bioinformatics. 2008;24(3):333–40. https://doi.org/10.1093/bioinformatics/btm604.
40. Shakhnovich EI, Gutin AM. Engineering of stable and fast-folding sequences of model proteins. Proc Nat Acad Sci. 1993;90(15):7195–9. https://doi.org/10.1073/pnas.90.15.7195.
41. Kuhlman B, Dantas G, Ireton GC, Varani G, Stoddard BL, Baker D. Design of a novel globular protein fold with atomic-level accuracy. Science. 2003;302(5649):1364–8. https://doi.org/10.1126/science.1089427.
42. Taylor WR, Sadowski MI. Structural constraints on the covariance matrix derived from multiple aligned protein sequences. PLOS One. 2011;6(12):28265. https://doi.org/10.1371/journal.pone.0028265.
43. Otter DW, Medina JR, Kalita JK. A survey of the usages of deep learning for natural language processing. IEEE Trans Neural Netw Learn Syst. 2021;32(2):604–24. https://doi.org/10.1109/TNNLS.2020.2979670.
44. Bahdanau D, Cho K, Bengio Y. Neural Machine Translation by Jointly Learning to Align and Translate. arXiv 2016. https://doi.org/10.48550/arXiv.1409.0473
45. Vig J, Madani A, Varshney LR, Xiong C, Socher R, Rajani NF. BERTology Meets Biology: Interpreting Attention in Protein Language Models. arXiv 2021. https://doi.org/10.48550/arXiv.2006.15222
46. Niu Z, Zhong G, Yu H. A review on the attention mechanism of deep learning. Neurocomputing. 2021;452:48–62. https://doi.org/10.1016/j.neucom.2021.03.091.
47. Ofer D, Brandes N, Linial M. The language of proteins: NLP, machine learning & protein sequences. Comput Struct Biotechnol J. 2021;19:1750–8. https://doi.org/10.1016/j.csbj.2021.03.022.
48. Barrat-Charlaix P, Muntoni AP, Shimagaki K, Weigt M, Zamponi F. Sparse generative modeling via parameter reduction of Boltzmann machines: application to protein-sequence families. Phys Rev E. 2021;104(2): 024407. https://doi.org/10.1103/PhysRevE.104.024407.
49. Gao C-Y, Zhou H-J, Aurell E. Correlation-compressed direct-coupling analysis. Phys Rev E. 2018;98(3): 032407. https://doi.org/10.1103/PhysRevE.98.032407.
50. Heinzinger M, Elnaggar A, Wang Y, Dallago C, Nechaev D, Matthes F, Rost B. Modeling aspects of the language of life through transfer-learning protein sequences. BMC Bioinf. 2019;20(1):723. https://doi.org/10.1186/s12859-019-3220-8.
51. Rao R, Meier J, Sercu T, Ovchinnikov S, Rives A. Transformer Protein Language Models Are Unsupervised Structure Learners. bioRxiv (2020). https://doi.org/10.1101/2020.12.15.422761
52. Rives A, Meier J, Sercu T, Goyal S, Lin Z, Liu J, Guo D, Ott M, Zitnick CL, Ma J, Fergus R. Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. Proc Nat Acad Sci. 2021;118(15):2016239118. https://doi.org/10.1073/pnas.2016239118.

Publisher's Note