Impact of phylogeny on structural contact inference from protein sequence data

Nicola Dietler$^{1,2}$, Umberto Lupo$^{1,2}$ and Anne-Florence Bitbol$^{1,2}$

$^1$Institute of Bioengineering, School of Life Sciences, Ecole Polytechnique Fédérale de Lausanne (EPFL), 1015 Lausanne, Switzerland
$^2$SIB Swiss Institute of Bioinformatics, 1015 Lausanne, Switzerland

Local and global inference methods have been developed to infer structural contacts from multiple sequence alignments of homologous proteins. They rely on correlations in amino acid usage at contacting sites. Because homologous proteins share a common ancestry, their sequences also feature phylogenetic correlations, which can impair contact inference. We investigate this effect by generating controlled synthetic data from a minimal model where the importance of contacts and of phylogeny can be tuned. We demonstrate that global inference methods, specifically Potts models, are more resilient to phylogenetic correlations than local methods, based on covariance or mutual information. This holds whether or not phylogenetic corrections are used, and may explain the success of global methods. We analyse the roles of selection strength and of phylogenetic relatedness. We show that sites that mutate early in the phylogeny yield false positive contacts. We consider natural data and realistic synthetic data, and our findings generalize to these cases. Our results highlight the impact of phylogeny on contact prediction from protein sequences and illustrate the interplay between the rich structure of biological data and inference.

1. Introduction

Statistical inference from biological data is currently making great advances, thanks to the rapid growth of available data, and to the development and application of inference methods. These methods range from interpretable models inspired by statistical physics to deep learning approaches. In this exciting context, it is important to understand the impact of the rich and complex structure of biological data on the performance of inference methods [1].

In particular, progress in sequencing has caused a spectacular growth of available genome sequences. This allows inference to be performed on large multiple sequence alignments (MSAs) of families of homologous proteins. Because proteins in the same family share the same ancestry, as well as similar three-dimensional structures and functional properties, these MSA data are highly structured. In particular, MSAs comprise correlations coming from phylogeny [2–4], as well as from structural and from functional constraints [5,6]. Pairwise maximum entropy models [7,8], known as Potts models or direct coupling analysis (DCA) [9], trained on natural MSAs, have revealed structural contacts [10–12]. They have also been employed to analyse mutational effects [13–16], protein evolution [17] and conformational changes [18,19], to design proteins [20], as well as to predict interaction partners among paralogs [21,22] and protein–protein interaction networks [23,24]. The key idea of these fruitful approaches is coevolution: amino acids that are in contact in the three-dimensional structure of proteins need to maintain physico-chemical complementarity, yielding correlations in amino acid usage at contacting sites. Recently, AlphaFold, a deep learning approach exploiting the breadth of available sequences and experimentally determined structures, has brought...
major advances to the computational prediction of protein structures from sequences [25]. Despite the important methodological differences with DCA, coevolution between amino acids is also an important ingredient of AlphaFold, which starts by constructing an MSA of homologues when given a protein sequence as input [25,26].

How does the complex structure of MSA data impact the performance of statistical inference? In MSAs, correlations due to phylogeny [2,4] and due to selection to preserve structure [10–12] and function coexist. Thus, phylogenetic correlations can impair the inference of structural contacts. This effect has been demonstrated both when using mutual information (MI) [27] and when using inferred Potts models [4,28–30]. It has motivated the development and use of various empirical corrections aimed at reducing the impact of phylogenetic correlations [27,31], such as the average product correction (APC) [32,33], phylogenetic reweighting [9–11,33,34] and nested coevolution [35].

How do phylogenetic correlations impact MSA properties and impair contact prediction using MI, covariance or inferred Potts models? To address this question, we generate synthetic data from a minimal model where the amount of correlations from structural contacts and from phylogeny can be fully controlled. We analyse the impact of the strength of natural selection to preserve structure, the impact of phylogenetic relatedness and their interplay. While phylogeny impairs contact prediction, we show that Potts models are more robust to phylogeny than local scores, even in the absence of explicit phylogenetic corrections, which are useful in both cases. This robustness to phylogeny comes on top of the ability of Potts models to disentangle direct and indirect correlations (which only matters for strong selection), and may explain the success of these methods. We further show that sites that mutate early in the phylogeny yield false positive (FP) contacts. Next, we analyse natural protein sequence data and realistic synthetic data generated from models inferred on natural data, either with or without phylogeny. Our findings from the minimal model generalise well to these cases.

2. Results

2.1. Minimal model with structural constraints and phylogeny

Separating correlations from structural contacts and from phylogeny is extremely tricky in natural data [36]. Thus, to assess the impact of phylogeny on structural contact prediction, we generate and study synthetic data with controlled amounts of correlations from structural constraints and from phylogeny. For this, we consider a minimal model, where protein sequences are represented by sequences of Ising spins, each spin sitting on one node of a fixed Erdős–Rényi random graph. Structural constraints, representing contacts in the three-dimensional protein structure, are modelled by pairwise couplings on the edges of this graph. We then sample independent equilibrium sequences using a Metropolis–Hastings algorithm (see Models and methods and electronic supplementary material, figure S1). This yields datasets of sequences that are only subject to the structural constraints defined by the graph. These datasets are regarded as MSAs. All correlations between sites then arise from these constraints or from finite-size effects. In order to introduce phylogeny in a controlled way, we start from one equilibrium sequence, considered as the ancestral sequence, and we evolve sequences on a binary branching tree with a fixed number $\mu$ of accepted mutations, i.e. substitutions, per branch. Mutations are accepted using the same Metropolis criterion as when generating independent equilibrium sequences. The sequences at the leaves of the tree yield an MSA that incorporates structural constraints and phylogeny. Both of them can yield correlations between the columns of the MSA. Our method to generate data is illustrated in figure 1, and described in detail in Models and methods. Note that the case with phylogeny and no selection is a limiting case of our model (see below).

2.2. Impact of phylogeny on contact prediction

What is the impact of phylogeny on contact prediction in this minimal model? Varying $\mu$ allows us to tune the amount of phylogeny. A small $\mu$ means that sequences are very closely related, while independent equilibrium sequences are recovered in the limit $\mu \to \infty$. In figure 2, we show the performance of contact prediction versus $\mu$. This performance is quantified by the TP fraction of predicted contacts, evaluated at the number of edges in the graph (see Models and methods), and we compare it with that obtained for equilibrium datasets. We perform contact inference with four different methods: two local ones, based on covariance and MI between sites, and two global ones, based on inferring DCA Potts models that fit the one- and two-body frequencies of the data. Local methods directly compare pairwise statistics, while global methods aim to infer a probability distribution for the whole sequence. The two global methods, mean-field DCA (mfDCA) [10,11] and pseudolikelihood maximization DCA (plmDCA) [32,37] (see Models and methods), employ different approximation schemes in the Potts model inference. In all cases, we present results both with and without the APC, which was introduced to correct for biases due to conservation and phylogeny when using MI to predict structural contacts [27]. This correction improves contact inference performance by DCA on natural protein sequence data [32,37], and is widely used. Furthermore, we find that it improves contact prediction performance substantially more than phylogenetic reweighting when using synthetic data generated from models inferred on natural data (see below). For equilibrium data, all inference methods yield a very good performance, with TP fractions close to 1. For data generated with phylogeny, the equilibrium results are recovered for large $\mu$, as expected. However, inference performance is substantially impaired for smaller values of $\mu$ (corresponding to strong phylogeny). Importantly, global (DCA) methods are more resilient to phylogenetic noise than local ones (covariance and MI) for small and intermediate values of $\mu$. This result is robust to whether APC is used or not, but we note that APC improves the performance of local methods when phylogeny is strong, while it has almost no effect on the performance of global methods. For instance, for $\mu = 15$, the TP fraction obtained by plmDCA is 28% (resp. 17%) higher than that obtained by MI without (resp. with) APC. APC generally improves the performance of both local [27] and global methods [32,37] on natural sequence data (see also our results below on natural and more realistic data). The success of APC with global methods may then be
2.3. Impact of selection strength on contact prediction without phylogeny

In addition to $\mu$, another important parameter of our model is the Monte Carlo sampling temperature $T$ (see figure 1 and equation (4.2)). It plays the role of an inverse selection strength, where selection aims to preserve structure. Note that other selective pressures are not considered in our minimal model. For large $T$, almost all mutations are accepted, and structural constraints are weak. Therefore, in the limit $T \to \infty$, selection vanishes and all correlations are from phylogeny. Conversely, for small $T$, essentially only the mutations that decrease the energy of the sequence of spins are accepted, and structural constraints become stringent. Figure 3 shows the performance of contact prediction versus $T$ for equilibrium sequences (no phylogeny), and for sequences generated with $\mu = 15$ and $\mu = 5$. At equilibrium, our model possesses a ferromagnetic-paramagnetic phase transition, whose approximate temperature $T_C = 4.2$ was found by inspecting magnetization histograms [38]. For equilibrium sequences at low $T$, inference performance without APC is poor (figure 3, left panel), and DCA methods perform substantially better than local methods (which are overlapping). The performance of DCA methods increases with $T$, and reaches a perfect score before $T_C$. Without APC, local methods also reach a perfect contact prediction score, but after $T_C$. However, they perform slightly better than DCA at higher temperatures. These trends are consistent with those previously described in [1,38], which analysed synthetic data generated independently at equilibrium, i.e. without phylogeny. Inference is difficult either if sequences are all frozen and redundant (very small $T$) or if sequences are too noisy (very large $T$), yielding better performance for intermediate values of $T$. Furthermore, APC substantially improves the performance of local methods at equilibrium, and makes their performance comparable to that of global methods even at small $T$ (figure 3, right panel). In electronic supplementary material, section S3, we show that the weak performance of local methods on equilibrium data at low $T$ and without APC is associated with conservation properties (see electronic supplementary material, figure S3), partly mitigated by regularization in global methods. Furthermore, in electronic supplementary material, section S4, we present a detailed comparison with [1], which shows that our results are fully consistent with theirs, and that regularization is crucial to the performance of DCA (see electronic supplementary material, figure S4).

2.4. Interplay between phylogeny and selection strength

What is the impact of phylogeny on the temperature dependence of contact prediction performance? As for equilibrium data, the best performance is obtained for intermediate values of $T$ with phylogenetic data (see figure 3). However, the maximum of performance shifts to lower values of $T$
when phylogeny is increased. This may be because, when phylogeny is strong, stringent selection (low $T$) keeps contacts relevant, thereby compensating for phylogeny, while at high $T$, the combination of thermal noise and phylogeny makes contact inference particularly tricky. Moreover, with phylogeny, DCA methods yield substantially better results than local methods for intermediate to high values of $T$, by contrast with the equilibrium case. This holds whether or not APC is used, even though APC improves the performance of local methods. Note, however, that MI slightly outperforms DCA methods at small values of $T$. Overall, the fact that DCA outperforms local methods in a wide range of selection strengths in the presence of phylogeny (generalizing the result observed at $T=5$ in figure 2) is much better in line with observations on natural data than with results obtained without phylogeny, where DCA yields at best marginal improvement compared with local methods with APC. Thus, robustness to phylogeny may be what makes DCA superior to local methods. We also note that plmDCA outperforms mfDCA when phylogeny is strongest ($\mu = 5$), which is reminiscent of the results obtained on natural data [32]. In electronic supplementary material, section S4, we consider a denser Erdős–Rényi graph, and our results are robust to this change (see electronic supplementary material, figures S4 and S5). However, larger values of pseudocounts or regularization strength than the usual ones are then helpful (see also electronic supplementary material, figure S6). The contact density used throughout and this denser one are both in the range observed in natural proteins (see electronic supplementary material, table S1).

2.5. Impact of graph properties on contact inference

Before exploring in more detail the impact of phylogeny, let us analyse the impact of graph properties, reflecting structural constraints, on equilibrium results. Indeed, in phylogenetic data, this structural signal needs to be disentangled from phylogenetic correlations. For each pair of sites, we consider the shortest path length $L$ in the graph connecting the two sites ($L = 1$ denotes contacting sites. Normalized histograms of coevolution scores for all pairs of sites are shown versus $L$ in electronic supplementary material, figure S7 without APC and in electronic supplementary material, figure S8 with APC. For $T = 3 < T_C$, without APC and especially for local methods, histograms of pairs in contact (l = 1) have substantial overlap with other histograms, especially with those for $L = 2$, showing that indirect correlations impair inference. However, APC strongly mitigates this issue, allowing local methods to perform almost as well as global ones, as seen in figure 3. Close to $T_C$, histograms of coevolution scores feature no overlap for DCA methods, leading to perfect inference. The small overlaps that still exist for local methods are resolved by using APC. Thus, APC enables local methods to disentangle direct and indirect correlations at equilibrium almost as well as global ones. At $T = 5 > T_C$, all methods yield perfect inference, as indirect correlations become weaker in the paramagnetic phase. Note also that plmDCA performs better than mfDCA at separating pairs with $L = 1$ and $L = 2$.

In addition to $L$, we consider another descriptor of graph structure, namely the number $N$ of nearest neighbours of each site. To assess the impact of $N$ on the variance of equilibrium data, we perform principal component analysis (PCA) focusing on the matrix of covariances between sequences [2]. Electronic supplementary material, figure S9 shows the coordinates of sites on the first principal component versus $N$. At all temperatures considered, and most strongly at low ones, these coordinates are correlated with $N$, showing that connectivity is a crucial ingredient of the variance of sites. Indeed, groups of connected spins tend to be aligned together at low temperatures, while isolated spins are independent. This difference is strongest at low $T$ (see equation (4.2)).

2.6. Origin and impact of phylogenetic correlations

How does phylogeny impair contact inference? What pairs of sites appear to be coupled due to phylogeny, while they are actually not coupled? To address these questions, we focus on the FP pairs which have the highest scores. Our synthetic data make it possible to investigate the states of these pairs of sites through the phylogenetic tree, starting from the ancestral sequence. The states of the top three FP pairs found by each method are shown in figure 4 for a dataset generated at $\mu = 5$ and $T = 5$ (i.e. with strong phylogeny). We observe large blocks of colours in our representation, meaning that these pairs of sites underwent mutations early in the phylogeny, leading to strongly correlated sites in the last generation. Indeed, early mutations at two sites produce correlations among them, as illustrated in figure 1b. This effect can be quantified by attributing to each pair of sites an index $G$, which is the first generation at which each of the spins of this pair has flipped (i.e. mutated), in at least one of the sequences generated thus far, with respect to its state in the ancestral sequence (see figure 1b). All top FP pairs in figure 4 have small $G$ values, and this effect is even stronger for local methods ($G = 1$ or 2) than for DCA methods ($G = 2$–4). Thus, pairs of sites mutating early in the phylogeny are more likely to have high coevolution
The indices of the sites as in figure 2, using the same contact map (Erdös–Rényi graph). The grey vertical line shows the approximate phase transition temperature $T_c$. All results are averaged over 100 realizations.

Figure 3. Impact of sampling temperature on contact prediction. The TP fraction is shown versus the Monte Carlo sampling temperature $T$ for four different inference methods (covariance, MI, mfDCA, plmDCA). The left panel shows results without APC, and the right panel with APC. Three cases are considered: equilibrium data (no phylogeny), and numbers of accepted mutations per branch $\mu = 15$ and $\mu = 5$ (increasing impact of phylogeny). Data are generated as in figure 2, using the same contact map (Erdös–Rényi graph). The grey vertical line shows the approximate phase transition temperature $T_c$. All results are averaged over 100 realizations.

Figure 4. Evolution of false positive (FP) pairs along the phylogenetic tree. For each inference method (MI, covariance, mfDCA, plmDCA), the state of the top three FP pairs of sites $(i, j)$, wrongly inferred as contacts from the evolved sequences, is depicted along the phylogenetic tree. This allows to track the history of these pairs $(i, j)$. At each generation, each vertical bar represents one sequence, and the colour of the bar indicates the state of the FP pair of focus within this sequence. Hence, the ancestor sequence is the single bar at generation 0, and the final generation (11) contains 2048 bars corresponding to all evolved sequences. Data are generated as in figure 2, using the same contact map (Erdös–Rényi graph), at $T = 5$ and $\mu = 5$ (strong phylogeny). Characteristics of each FP pair $(i, j)$ are reported, namely the indices of the sites $i$ and $j$, their number $N$ of neighbours in the graph (i.e. sites in contact with $i$ or $j$), the length $L$ of the shortest path connecting $i$ to $j$ in the graph ($L = 1$ for contacts), and the earliest mutation generation $G$ where both $i$ and $j$ have mutated with respect to the ancestral sequence.

Figure 5 shows the distribution of coevolution scores versus the index $G$ as a violin plot (i.e. Gaussian kernel-smoothed vertical histograms). It sheds light on an overall dependence of the coevolution scores on $G$, with larger coevolution scores observed for small $G$, i.e. early in the phylogenetic history.

The shortest path length $L$ is also indicative on the phylogenetic origin of their high coevolution scores. Indeed, $L > 2$ for all of them but one, while if indirect correlations were the main cause of these FPs, we would expect more occurrences of $L=2$. Note that, in [39], we showed that these pairs of sites are particularly useful for the inference of interacting partners among the paralogs of two protein families, where phylogenetic correlations are useful.

How do these observations made on a few top FPs generalize? To assess this, we compute the index $G$ of all pairs of sites $(i, j)$—including both contacts and non-contacts—in multiple synthetic datasets generated at $\mu = 5$ and $T = 5$. (Note that the results that follow also hold when restricting to all FPs or to all non-contacts, which are much more numerous than contacts.) We then group pairs of sites that have a given value of $G$. Figure 5 shows the distribution of coevolution scores versus the index $G$ as a violin plot (i.e. Gaussian kernel-smoothed vertical histograms). It sheds light on an overall dependence of the coevolution scores on $G$, with larger coevolution scores observed for small $G$, i.e. early
mutations. Moreover, this dependence is weaker for global methods than for local ones, and among global methods, it is weaker for plmDCA than for mfDCA. These conclusions hold both without and with APC. This confirms that DCA, especially plmDCA, is more robust to phylogenetic correlations than local inference methods, which could partly explain its success on natural protein sequences. The behaviour of the median score versus \( G \) (red curve in figure 5) is further investigated in electronic supplementary material, figure S10. Without APC, the median decays abruptly with \( G \), especially for local methods. This decrease is less abrupt and closer to linear for plmDCA, confirming its lower sensitivity to the phylogeny. APC substantially reduces the median for small \( G \) for all methods, thereby mitigating the impact of phylogeny on inference performance.

Besides, electronic supplementary material, figure S11 shows the average conservation of sites involved in a pair versus \( G \). It shows that pairs of sites undergoing mutations early in the phylogeny are often weakly conserved. Qualitatively, sites with early mutations will often be in different states (up and down) in sequences at the leaves of the tree, yielding a poor conservation but a strong correlation. This point could be exploited to disentangle TP and FP pairs when the underlying phylogeny is not known. The inferred fields could also be useful to this end. However, we expect the general case to be more complex as conservation can arise both from single-site constraints modelled by non-zero fields and from late divergences along the phylogeny. Note that in sector analysis, the focus is on conserved mutations will often be in different states (up and down) in sequences at the leaves of the tree, yielding a poor conservation but a strong correlation. This point could be exploited to disentangle TP and FP pairs when the underlying phylogeny is not known. The inferred fields could also be useful to this end. However, we expect the general case to be more complex as conservation can arise both from single-site constraints modelled by non-zero fields and from late divergences along the phylogeny. Note that in sector analysis, the focus is on conserved mutations.

Figure 5. Impact of phylogeny on coevolution scores. Violin plots of the coevolution scores of all pairs of sites \((i,j)\), inferred by four different methods (MI, covariance, mfDCA, plmDCA), are shown versus the earliest generation \( G \) where both \( i \) and \( j \) have mutated with respect to the ancestral sequence. The left panels show the raw Frobenius norm scores (absolute value here) and the right panels show these scores corrected with APC. Data are generated as in figure 2, using the same contact map (Erdös–Rényi graph), at \( T = 5 \) and \( \mu = 5 \). Couplings were inferred on 100 datasets of 2048 sequences each, and aggregated.

2.7. Extension to more realistic synthetic data, generated from models inferred on natural data

How well does the insight gained from our minimal model apply to natural data? To address this question, we generate synthetic data directly using DCA Potts models inferred on natural MSAs, and phylogenies inferred on the same MSAs. Specifically, we construct three different datasets for each of the four protein families in electronic supplementary material, table S1: natural sequences, equilibrium sequences from the inferred DCA model, and sequences generated using the same inferred DCA model, but along an inferred phylogenetic tree (see Models and methods). These synthetic datasets are closer to natural data than those generated from our minimal model, but retain the advantage that we know when in the phylogeny each mutation occurred. Figure 6 shows the effect of phylogenetic correlations on couplings inferred by MI and plmDCA with or without APC on this data. The main conclusion from our minimal model still holds: higher coevolution scores are obtained for pairs of sites that mutate early in the phylogeny, for both MI and plmDCA without APC. The magnitude of this trend depends on the protein family considered, which may be due to the diversity of their phylogenies (note that the ratio of the effective number of sequences to the actual number of sequences takes diverse values among these families, see electronic supplementary material, table S1). How do phylogenetic corrections impact this trend? We first compare the effects of different phylogenetic corrections (APC, phylogenetic reweighting) on the performance of contact inference on natural sequences (see electronic supplementary material, table S2). While both corrections enhance performance and their combination is best, APC improves contact prediction substantially more than phylogenetic reweighting.

Consistently, figure 6 shows that APC strongly impacts the dependence of coupling strength on phylogeny, as we then generally obtain smaller absolute Pearson correlations of the coevolution scores with the earliest mutation, especially for plmDCA. The overall trend is even reversed by APC for MI, as APC-corrected MI scores are lower for pairs of sites that mutate early in the phylogeny.

The contact maps for each of these natural and realistic synthetic datasets are shown in electronic supplementary material, figures S12 and S13, using MI and plmDCA,
respectively, and the TP fractions with or without phylogenetic corrections (phylogenetic reweighting and/or APC) are shown in electronic supplementary material, table S3. Note that here, contrary to our minimal model, the couplings of the inferred DCA models contain some phylogenetic contributions. However, it is fair to compare the equilibrium (independent) and phylogenetic datasets, using the contact map from the bmDCA model (see Models and methods and [41]) as ground truth. We find that phylogeny deteriorates contact prediction performance in these datasets, corroborating the findings from our minimal model. Besides, electronic supplementary material, table S3 confirms that phylogenetic corrections almost always improve the performance of contact prediction. We also find that the number of FP predicted contacts having shortest path length \( L = 2 \), i.e. indirect correlations, is smaller for plmDCA than for MI, consistently with electronic supplementary material, figure S7 (see electronic supplementary material, table S3). Finally, and perhaps most interestingly, the number of these FPs with \( L = 2 \) is often smaller in the dataset generated with phylogeny than in the equilibrium one, even though the TP fractions decrease as well, suggesting that most FPs are due to phylogeny (and not to the network of contacts) in the phylogenetic datasets.

### 3. Discussion

Global statistical methods, known as Potts models or DCA, outperform local methods based on covariance or MI at the task of unsupervised prediction of structural contacts from protein sequence data [9–11]. Perhaps surprisingly, these global methods do not outperform local methods as clearly when applied to synthetic data generated independently at equilibrium from DCA models [1]. A usual justification for the success of global methods for protein contact prediction is that they allow to disentangle direct and indirect correlations in the data [9], an effect that is present in synthetic data as well, and can be tuned by varying sampling temperature around a ferromagnetic-paramagnetic phase transition [1,38]. An important difference between natural data and synthetic data sampled independently at equilibrium is the fact that homologous natural sequences share the same ancestry, and thus feature correlations due to phylogeny [2,4]. These correlations obscure structural ones in the identification of structural contacts both by local [27] and global statistical methods [4,28–30].

In this context, we investigated the impact of phylogeny on inference performance by generating synthetic data from a minimal model incorporating both structural contacts and phylogeny. We showed that substantial correlations appear between sites that mutate early in the phylogeny, yielding FP contacts. We found that Potts models are substantially more robust to phylogeny than local methods. This result holds whether or not phylogenetic corrections are used. This robustness to phylogeny contributes to explaining the success of global methods. We showed that our findings from the minimal model generalize well to more realistic synthetic data generated from models inferred on natural data.

The resilience of Potts models to phylogenetic correlations can be understood qualitatively in the simple mean-field approximation, where the DCA couplings are the elements of the inverse covariance matrix of the MSA. Indeed, phylogenetic correlations are associated with large eigenvalues of the covariance matrix [2,4], whose overall impact is reduced by inverting the covariance matrix [4]. Our findings illustrate the crucial impact of data structure on inference performance. Protein MSAs are highly structured datasets, with correlations coming from functional constraints such as structural contacts, as well as from

![Figure 6. Impact of phylogeny on coevolution scores for more realistic data. Violin plots of the coevolution scores of all pairs of sites \((i, j)\), inferred by MI or plmDCA, without or with APC, are shown versus the earliest mutation \( G' \) where both \( i \) and \( j \) have mutated with respect to the ancestral sequence. To determine \( G' \), we count the number \( G_j \) (resp. \( G_i \)) of mutations undergone by a sequence when following the branches of the phylogenetic tree from the ancestral sequence to the first one where \( i \) (resp. \( j \)) has mutated, and compute \( G = \max (G_j, G_i) \). Pairs \((i, j)\) are binned by the value of \( G' \) for the violin plot representation. Two Pearson correlation coefficients are shown: the first one, \( \rho_s \), is obtained when considering all coupling values, and the second one, \( \rho_{hs} \), is obtained when considering only the median of the coupling values. Note that no binning on \( G' \) is employed in the computation of these Pearson correlations. Couplings were inferred on 100 realizations of the data generation in each case, and aggregated together.](https://royalsocietypublishing.org/journal/rsif)
4. Models and methods

4.1. Synthetic data generation

4.1.1. General approach

We generate synthetic sequences with fixed structural constraints, modeling those that would exist in a given protein family that needs to maintain structure and function. We consider two types of datasets. The first one is generated at equilibrium under these structural constraints. The second one has an evolutionary history given by a simple phylogeny, and we assume that the same structural constraints have existed throughout this evolutionary history.

To model structural constraints, we use an Erdős-Rényi random graph with 200 nodes and edge probability $q = 0.02$. This graph is held fixed because it models the structural constraints of a given protein family. Each node corresponds then to an amino acid site, and each edge to a structural contact. This number of sites and this density of contacts are in the range of those observed in natural proteins (see electronic supplementary material, table S1). However, because contact density depends on the threshold used to define contacts, it is interesting to consider larger values of $q$ (see electronic supplementary material, table S1). In electronic supplementary material, section S4, we also consider a graph with $q = 0.2$ [1]. For simplicity, we model amino acids as Ising spins, i.e., binary variables. We consider that couplings between sites exist on the graph edges, and within our minimal model, binary variables. We consider that couplings between sites $i$ and $j$ ($J_{ij} = 1$ if $(i, j) \in \mathcal{E}$, and 0 otherwise).

4.1.2. Generating sequences with structural constraints only

Independent equilibrium sequences are generated by Metropolis Monte Carlo sampling using the Hamiltonian in equation (4.1), which models structural constraints. Each sequence is randomly and independently initialized. Then, moves are proposed, using one of two different methods, either by flipping single spins or clusters of spins (Wolff cluster algorithm) [46]. In both methods, these moves are accepted or rejected according to the Metropolis criterion, with probability

$$p = \min \left[ 1, \exp \left( -\frac{\Delta H}{T} \right) \right],$$

where $\Delta H$ is the difference in the value of $H$ after and before the flip, and $T$ is the Monte Carlo sampling temperature. This process is iterated until a certain number of moves are accepted. We choose this number so that equilibrium is reached, which can be checked by the saturation of the absolute magnetization (see electronic supplementary material, figure S1). Note that the decorrelation time of absolute magnetization is similar to the magnetization saturation time.

4.1.3. Generating sequences with structural constraints and phylogeny

To include phylogeny in the sequence generation process, we start from an equilibrium ancestor sequence produced as explained above at a given temperature $T$. Then, this sequence is evolved by successive duplication and mutation events (‘generations’) on a binary branching tree, for a total of $G = 11$ generations (see figure 1). Mutations are modelled as single spin flips, and each of them is accepted according to the Metropolis criterion at temperature $T$ (see equation (4.2)) with the Hamiltonian in equation (4.1) accounting for structural constraints. We perform a fixed number $\mu$ of accepted mutations on each branch of the phylogenetic tree. The sequences at the leaves of the phylogenetic tree provide a dataset of $2^G = 2^{11} = 2048$ sequences.

Because we start from an ancestral equilibrium sequence, and then employ the Metropolis criterion at the same temperature $T$, all sequences in the phylogeny are equilibrium sequences. Thus, these sequences comprise correlations from the structural constraints, as without phylogeny. Their relatedness adds extra correlations that we call phylogenetic correlations.

4.1.4. Generating more realistic sequences

To extend our study to more realistic data, we generated sequences from Potts models inferred on natural sequence data. Specifically, natural sequences (Pfam Full MSAs [47]) were retrieved for four different protein families (PF00072, PF00512, PF00595, PF02518). Next, a bmDCA Potts model [41] was inferred on each of these natural MSAs. Indeed, bmDCA has a good generative power, demonstrated experimentally [20], and thus allows to generate new sequences by Metropolis MCMC sampling [41]. (Adaptive cluster expansion, ACE [48,49], an alternative generative method, proved too computationally demanding for our analysis.) We used this procedure to produce an equilibrium dataset.
of sequences from each of the inferred Potts models. We also generated another set of sequences that contains phylogeny using each of these Potts models [39,43]. For this, we first inferred a phylogenetic tree using FastTree [50] from the natural sequences. At its (arbitrary) root, we place a sequence generated at equilibrium by bmDCA. Then, the sequence evolves according to the inferred phylogeny, where proposed mutations are accepted using the same Metropolis criterion as for the equilibrium data. All sequences at the leaves are collected to form the phylogenetic dataset. Thus, three datasets are available for each protein family: a natural dataset, an equilibrium synthetic dataset and a phylogenetic synthetic dataset.

4.2. Inferring contacts from sequences

4.2.1. General approach

We employ methods developed for contact prediction in natural proteins to infer the edges (or contacts) in the graph, and investigate how they are affected by phylogeny. In this spirit, the sequences (either a set of equilibrium sequences generated at a given \( T \), or all sequences at the leaves of a given phylogenetic tree) are put together to form a MSA, i.e. a matrix where each row is a sequence and each column is a site. Note that here, there is no alignment issue because each node of the graph is well identified. We compare four different inference methods, which attribute a score to each pair of sites (nodes) in the Erdős–Rényi graph. Top-scoring pairs are predicted contacts. The first two inference methods are local methods, based, respectively, on covariance (C) and on MI [27]. The other two methods are global methods, which consist in learning a maximum entropy model consistent with the one- and two-site frequencies observed in the data. These pairwise maximum entropy models, also known as Potts models, or DCA [9], can be approximately inferred in various ways. We consider two widely used versions, namely mean-field DCA (mfDCA) [10,11], which is the simplest one, and plmDCA [32,37], which is the state-of-the-art contact prediction method based on Potts models.

All these methods start from the single-site frequencies of each state \( \sigma_i \) at each site (or column in the MSA) \( i \in \{1, \ldots, \ell \} \), denoted by \( \hat{f}_i(\sigma_i) \), and the two-site frequencies \( \hat{f}_{ij}(\sigma_i, \sigma_j) \). In addition, MI and DCA methods employ regularization schemes. First, a pseudocount is added in the computation of the frequencies for MI [44,51] and for mfDCA [10,11], preventing divergences when computing MI or when inverting the covariance matrix in mfDCA. The pseudocount-corrected frequencies read

\[
\hat{f}_i(\sigma_i) = \frac{\lambda}{2} + (1 - \lambda)\hat{f}_i(\sigma_i), \tag{4.3}
\]

\[
\hat{f}_{ij}(\sigma_i, \sigma_j) = \frac{\lambda}{4} + (1 - \lambda)\hat{f}_{ij}(\sigma_i, \sigma_j) \quad \text{for } i \neq j \tag{4.4}
\]

and

\[
\hat{f}_{ii}(\sigma_i, \sigma_i) = \delta_{\sigma_i, \sigma_i} \hat{f}_i(\sigma_i), \tag{4.5}
\]

where \( \delta_{\sigma_i, \sigma_j} = 1 \) for \( \sigma_i = \sigma_j \) and 0 otherwise. This correction improves contact prediction from protein sequences by mfDCA [10,11]. In plmDCA, no pseudocount is employed, but an \( L^2 \) norm regularization is used, with strengths \( \lambda_h \) on fields and \( \lambda_f \) on couplings [32,37]. Unless otherwise specified, we use the standard values for all these regularization parameters, namely \( \lambda = 0.01 \) for MI, \( \lambda = 0.5 \) for mfDCA and \( \lambda_f = \lambda_h = 0.01 \) for plmDCA. An analysis of the impact of pseudocount and regularization strength on contact prediction by mfDCA and plmDCA is provided in electronic supplementary material, figure S6. We compare results with and without APC [27], but we do not employ phylogenetic reweighting [9] in the analysis of our minimal model, because the phylogeny is balanced in that case, which would yield threshold effects in this correction. Besides, we find in our analysis of natural and more realistic data that APC yields a more substantial performance increase than phylogenetic reweighting.

4.2.2. Covariance

The covariance between two sites \( i \) and \( j \) reads \( C_{ij} = \langle \sigma_i \sigma_j \rangle - \langle \sigma_i \rangle \langle \sigma_j \rangle \), where \( \langle \cdot \rangle \) denotes a mean across all sequences of the MSA. While this standard definition can be employed directly with Ising spins, the state of a site in a protein sequence is a categorical random variable with 21 possible states (the 20 natural amino acids and the gap), and therefore frequencies are generally employed instead of means. The two descriptions are equivalent for Ising spins. Indeed,

\[
\langle \sigma_i \rangle = \hat{f}_i(1) - \hat{f}_i(-1) = 2\hat{f}_i(1) - 1 \tag{4.6}
\]

and

\[
\langle \sigma_i \sigma_j \rangle = \hat{f}_{ij}(1, 1) - \hat{f}_{ij}(-1, 1) - \hat{f}_{ij}(-1, -1) + \hat{f}_{ij}(1, -1)
\]

\[= 1 - 2\hat{f}_{ij}(1) - 2\hat{f}_{ij}(1) + 4\hat{f}_{ij}(1, 1), \tag{4.7}\]

where we have employed the marginalization relationships, \( \hat{f}_i(\sigma) = \hat{f}_{i1}(\sigma, 1) + \hat{f}_{i0}(\sigma, -1) \) and \( \hat{f}_{ij}(\sigma) = \hat{f}_{ij1}(\sigma, 1) + \hat{f}_{ij0}(\sigma, -1) \). \( \hat{f}_{ij}(1) + \hat{f}_{ij}(-1) = 1 \), which yield \( \hat{f}_{ij}(1) + \hat{f}_{ij}(-1) = 1 \), and the normalization relationship \( \hat{f}_{ij}(1) + \hat{f}_{ij}(-1) = 1 \), which yield \( \hat{f}_{ij}(1) + \hat{f}_{ij}(-1) = 1 \), as well as \( \hat{f}_{ij}(-1, -1) = \hat{f}_{ij}(1, 1) \). Combining equations (4.6) and (4.7) yields

\[
C_{ij} = \langle \sigma_i \sigma_j \rangle - \langle \sigma_i \rangle \langle \sigma_j \rangle = 4\hat{f}_{ij}(1, 1) - \hat{f}_{ij}(1)\hat{f}_{ij}(1). \tag{4.8}\]

We employ the absolute value \( |C_{ij}| \) of the covariance to score each pair of sites \((i,j)\).

4.2.3. Mutual information

The MI of each pair of sites \((i,j)\) is computed as

\[
\text{MI}_{ij} = \sum_{\sigma_i, \sigma_j} \hat{f}_{ij}(\sigma_i, \sigma_j) \log \frac{\hat{f}_{ij}(\sigma_i, \sigma_j)}{\hat{f}_i(\sigma_i) \hat{f}_j(\sigma_j)} \tag{4.9}
\]

where \( \hat{f}_{ij}(\sigma_i, \sigma_j) \) and \( \hat{f}_i(\sigma_i) \) are the pseudocount-corrected one- and two-body frequencies (see equations (4.3)-(4.5)). Note that we use frequencies instead of probabilities to estimate MI, and do not correct for finite-size effects [52]. Indeed, we only compare scores computed on datasets with a given size, affected by the same finite-size effects.

4.2.4. Potts models (DCA)

DCA methods aim to infer the pairwise maximum entropy probability distribution of the data, which is the least constrained distribution consistent with the empirically measured one- and two-body frequencies. The probability of a sequence \( \sigma \) has the following form:

\[
P(\sigma) = \frac{\exp[-H(\sigma)]}{Z}, \tag{4.10}\]

where the Hamiltonian \( H \) is given by equation (4.1), i.e. the Potts model Hamiltonian, and \( Z \) is a normalization constant.
(partition function). The mDCA method provides a simple approximation of the couplings \( J_{ij} \) in the zero-sum (Ising) gauge \([10,11]\) given by \( J_{ij} = -\langle C \rangle_{ij} \) where \( C \) is the covariance matrix computed with the pseudocount-corrected frequencies (see equations (4.3)-(4.5)), whose elements read
\[
C_{ij} = (1 - \lambda)C_{ij} + \lambda\langle \sigma_i \sigma_j \rangle = (1 - \lambda)[\langle \sigma_i \sigma_j \rangle - (1 - \lambda)^2 \langle \sigma_i \rangle \langle \sigma_j \rangle]
\]
for \( i \neq j \) and \( C_{ii} = (1 - \lambda)^2C_{ii} + \lambda(2 - \lambda) - (1 - \lambda)^2 - \lambda(2 - \lambda) \). Finally, the plmDCA method provides another estimation \( J_{ij} \) of the couplings \( J_{ij} \) by maximizing the pseudolikelihood of equation (4.10). We also use the zero-sum gauge for plmDCA \([32,37]\), and we employ the absolute value \( |J_{ij}| \) of the inferred coupling to score each pair of sites \((i, j)\).

With natural sequences and realistic synthetic sequences with 21 states, the Frobenius norm of the \( 21 \times 21 \) matrix of coupling values \( J_{ij}(\alpha, \beta) \) is used to obtain one score per pair of sites \([10,11]\) and it is denoted by \( ||J|| \).

### 4.2.5. Evaluating performance

We assess performance via the fraction of correctly identified contacts among the \( N_{\text{contacts}} \) top-scoring pairs of sites, where \( N_{\text{contacts}} = 413 \) is the number of actual contacts, i.e. of edges in the Erdös–Rényi random graph used to generate the data. We refer to this fraction as the true positive fraction (TP fraction). Because we evaluate it for a number of predicted contacts equal to the number of actual contacts, we have TP + FP = TP + FN = \( N_{\text{contacts}} \), where TP denotes the true positives, FP is the false positives and FN is the false negatives. Thus, the TP fraction is equal to the sensitivity, recall, hit rate or true positive rate \( TPR = TP/(TP + FN) \), but also to the TP fraction is equal to the number of actual contacts, where TP denotes the true positives, FP is the false positives and FN is the false negatives. Thus, the TP fraction is equal to the sensitivity, recall, hit rate or true positive rate \( TPR = TP/(TP + FN) \), but also to the precision or positive predictive value \( PPV = TP/(TP + FP) \).

We employ the TP fraction (or PPV) as a measure of performance because the usual practice for natural sequence data is to predict the top-scoring pairs of sites as contacts. It is also interesting to evaluate how the quality of predictions depends on the number of predicted contacts. For this, one can consider the area under the receiver operating characteristic (AUC). We found very similar conclusions for the AUC as for the TP fraction (see electronic supplementary material, figures S4 and S5).

### 4.2.6. Extension to more realistic data

Inference on natural and more realistic data is performed using one local method and one global method, respectively, MI and plmDCA. For MI, we employ equation (4.9), using a pseudocount of 0.01, with \( \sigma_i \in \{1, \ldots, 21\} \). For plmDCA, we use the implementation of the PlmDCA package (https://github.com/pagnani/PlmDCA). Unless otherwise specified, the default values of the parameters of this package are used.

For natural and more realistic data, the number of predicted contacts is taken as \( N_{\text{pred}} = 2\epsilon \), where \( \epsilon \) is the length of the protein (number of amino acid sites, reported in electronic supplementary material, table S1), and performance is measured as the PPV (or TP fraction) at this number of predicted contacts \([10,11]\).

### Data accessibility

All relevant data are included in the manuscript or in the electronic supplementary material \([53]\). Our code is freely available at https://zenodo.org/record/7503931.

### Authors’ contributions

N.D.: conceptualization, investigation, methodology, software, visualization, writing—original draft, writing—review and editing; U.L.: investigation, resources, software, writing—review and editing; A.-F.B.: conceptualization, funding acquisition, investigation, supervision, writing—original draft, writing—review and editing.

All authors gave final approval for publication and agreed to be held accountable for the work performed therein.

### Conflict of interest declaration

We declare we have no competing interests.

### Funding

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

### Acknowledgements

N.D. and A.-F.B. thank Matthieu Wyart for helpful discussions.

### References

1. Ngampruetikorn V, Sachdeva V, Torrence J, Humplik J, Schwab DJ, Palmer SE. 2022 Inferring couplings in networks across order-disorder phase transitions. Phys. Rev. Res. 4, 023240. (doi:10.1103/PhysRevResearch.4.023240)
2. Casari G, Sander C, Valencia A. 1995 A method to predict functional residues in proteins. Nat. Struct. Biol. 2, 171–178. (doi:10.1038/nsb0295-171)
3. Halabi N, Rivoire O, Leibler S, Ranganathan R. 2009 Protein structures: evolutionary units of threedimensional structure. Cell 138, 774–786. (doi:10.1016/j.cell.2009.07.038)
4. Qin C, Colwell LJ. 2018 Power law tails in phylogenetic systems. Proc. Natl Acad. Sci. USA 115, 690–695. (doi:10.1073/pnas.1711913115)
5. Gobel U, Sander C, Schneider R, Valencia A. 1994 Correlated mutations and residue contacts in proteins. Proteins 18, 309–317. (doi:10.1002/prot.340180402)
6. Pazos F, Helmer-Citterich M, Ausiello G, Valencia A. 1997 Correlated mutations contain information about protein–protein interaction. J. Mol. Biol. 271, 511–523. (doi:10.1006/jmbi.1997.1198)
7. Jaynes ET. 1957 Information theory and statistical mechanics. Phys. Rev. 106, 620–630. (doi:10.1103/PhysRev.106.620)
8. Lapides AS, Giaud BG, Liu L, Stormo GD. 1999 Correlated mutations in models of protein sequences: phylogenetic and structural effects. In Proc. of the IMS/AMS Int. Conf. on Statistics in Molecular Biology and Genetics. IMS Lecture Notes — Monograph Series, vol. 33, pp. 236–256. Hayward, CA: Institute of Mathematical, Statistics and American Mathematical Society. (doi:10.2172/296863)
9. Weigt M, White RA, Szmuntant H, Hoch IA, Hwa T. 2009 Identification of direct residue contacts in protein–protein interaction by message passing. Proc. Natl Acad. Sci. USA 106, 67–72. (doi:10.1073/pnas.0800592106)
10. Marks DS, Colwell LJ, Sheridan R, Hofp TA, Pagnani A, Zecchina R, Sander C. 2011 Protein 3D structure computed from evolutionary sequence variation. PloS ONE 6, e28766. (doi:10.1371/journal.pone.0028766)
11. Morcos F et al. 2011 Direct-coupling analysis of residue coevolution captures native contacts across many protein families. Proc. Natl Acad. Sci. USA 108, E1293–1301. (doi:10.1073/pnas.1111471108)
12. Sulkowska JI, Morcos F, Weigt M, Hwa T, Onuchic JN. 2012 Genomics-aided structure prediction. Proc. Natl Acad. Sci. USA 109, 10 340–10 345. (doi:10.1073/pnas.1207864109)
13. Dwyer RS, Rici DP, Colwell LJ, Slaby TJ, Wingreen NS. 2013 Predicting functionally informative mutations in Escherichia coli BamA using evolutionary covariance analysis. Genetics 195, 443–455. (doi:10.1534/genetics.113.155861)
14. Cheng RR, Morcos F, Levine H, Onuchic JN. 2014 Toward rationally designing bacterial two-component signaling systems using coevolutionary information. Proc. Natl Acad. Sci. USA 111, E563–E571. (doi:10.1073/pnas.1322373111)
15. Cheng RR, Nordesjo O, Hayes RL, Levine H, Flores SC, Onuchic JN, Morcos F. 2016 Connecting the
sequence-space of bacterial signaling proteins to phenotypes using coevolutionary landscapes. Mol. Biol. Evol. 33, 3054–3064. (doi:10.1093/molbev/msv188)

16. Figliuzzi M, Jacquier H, Schug A, Tenaillon O, Weigt M. 2016 Coevolutionary landscape inference and the context-dependence of mutations in beta-lactamase TEM-1. Mol. Biol. Evol. 33, 268–280. (doi:10.1093/molbev/msv211)

17. de la Paz JA, Nartry CM, Yuvaraj M, Morcos F. 2020 Epistatic contributions promote the unification of incompatible models of neutral molecular evolution. Proc. Natl Acad. Sci. USA 117, 5873–5882. (doi:10.1073/pnas.1913071117)

18. Morcos F, Jana B, Hwa T, Onuchic JN. 2013 Evolutionary coupling analyses. Entropy 21, 1090. (doi:10.3390/e21110109)

19. Malinverni D, Marsili S, Barducci A, De Los Rios P. 2015 Large-scale conformational transitions and dimerization are encoded in the amino-acid sequences of Hsp70 chaperones. PLoS Comput. Biol. 11, e1004262. (doi:10.1371/journal.pcbi.1004262)

20. Russ WP et al. 2020 An evolution-based model for designing chioromate mutase enzymes. Science 369, 440–445. (doi:10.1126/science.aba3304)

21. Bitbol A-F, Dewey RS, Colwell LJ, Wingreen NS. 2016 Inferring interaction partners from protein sequences. Proc. Natl Acad. Sci. USA 113, 12180–12185. (doi:10.1073/pnas.1606762113)

22. Gueudre T, Baldassi C, Zamparo M, Weigt M, Pagani A. 2016 Simultaneous identification of specifically interacting paralogs and interpret protein contacts by direct coupling analysis. Proc. Natl Acad. Sci. USA 113, 12186–12191. (doi:10.1073/pnas.1605710718)

23. Cong Q, Anischchenko I, Orchinikov S, Baker D. 2019 Protein interaction networks revealed by proteome coevolution. Science 365, 185–189. (doi:10.1126/science.aav6718)

24. Green AG, Elbahaisy H, Brock KP, Maddamsetti R, Kohlhauser O, Marks DS. 2021 Large-scale discovery of protein interactions at residue resolution using co-evolution calculated from genomic sequences. Nat. Commun. 12, 1396. (doi:10.1038/s41467-021-23637-3)

25. Jumper J et al. 2021 Highly accurate protein structure prediction with AlphaFold. Nature 596, 583–589. (doi:10.1038/s41586-021-03819-2)

26. Roney JP, Orchinikov S. 2022 State-of-the-art estimation of protein model accuracy using AlphaFold. bioRxiv. (doi:10.1101/2022.03.11.484043)

27. Dunn SD, Wahl LM, Gloor GB. 2008 Mutual information without the influence of phylogeny or entropy dramatically improves residue contact prediction. Bioinformatics 24, 333–340. (doi:10.1093/bioinformatics/btn604)

28. Vorberg S, Seemayer S, Söding J. 2018 Synthetic protein alignments by CCmgen quantify noise in residue-residue contact prediction. PLoS Comput. Biol. 14, 1–25. (doi:10.1371/journal.pcbi.1006526)

29. Rodriguez Horta E, Barrat-Charlaix P, Weigt M. 2019 Toward inferring Potts models for phylogenetically correlated sequence data. Entropy 21, 1090. (doi:10.3390/e21110109)

30. Rodriguez Horta E, Weigt M. 2021 On the effect of phylogenetic correlations in coevolution-based contact prediction in proteins. PLoS Comput. Biol. 17, e1008957. (doi:10.1371/journal.pcbi.1008957)

31. Lichtarge O, Bourne HR, Cohen FE. 1996 An evolutionary trace method defines binding surfaces common to protein families. J. Mol. Biol. 257, 342–358. (doi:10.1006/jmbi.1996.0767)

32. Ekeberg M, Lovkvist C, Lan Y, Weigt M, Aurell E. 2013 Improved contact prediction in proteins: using pseudolikelihoods to infer Potts models. Phys. Rev. E 87, 012702. (doi:10.1103/PhysRevE.87.012707)

33. Hockenberg AJ, Wilke CO. 2019 Phylogenetic weighting does little to improve the accuracy of evolutionary coupling analyses. Entropy 21, 1000. (doi:10.3390/e21111000)

34. Malinverni D, Barducci A. 2020 Coevolutionary analysis of protein subfamilies by sequence reweighting. Entropy 21, 1127. (doi:10.3390/e21111127)

35. Colavin A, Atolia E, Bitbol A-F, Huang KC. 2022 Extracting phylogenetic dimensions of coevolution reveals hidden functional signals. Sci. Rep. 12, 820. (doi:10.1038/s41598-021-04269-1)

36. Weinstein EN, Amin AN, Frazer J, Marks DS. 2022 Non-identifiability and the blessings of misspecification in models of molecular fitness and phylogeny. bioRxiv. (doi:10.1101/2022.01.29.478324)

37. Ekeberg M, Hartonen T, Aurell E. 2014 Fast pseudolikelihood maximization for direct-coupling analysis of protein structure from many homologous amino-acid sequences. J. Comput. Phys. 276, 341–356. (doi:10.1016/j.jcp.2014.07.024)

38. Gandanilla-Perez CA, Barrat-Charlaix P, Weigt M. 2022 Combining phylogeny and coevolution improves the inference of interaction partners among paralogous proteins. (http://arxiv.org/abs/2208.11626)

39. Krauth W 2006 Statistical mechanics: algorithms and computations. Oxford Master Series in Physics. Oxford University Press.

40. Misty J et al. 2020 Pfam: the protein families database in 2021. Nucleic Acids Res. 49, 412–419. (doi:10.1093/nar/gkaa913)

41. Barton JP, De Leonardis E, Coucke A, Cocca S. 2016 ACE: adaptive cluster expansion for maximum entropy graphical model inference. Bioinformatics 32, 3089–3097. (doi:10.1093/bioinformatics/btw328)

42. Bravi B, Ravasio R, Brito C, Wyatt M. 2020 Direct coupling analysis of epitasis in allosteric molecules. PLoS Comput. Biol. 16, e1007630. (doi:10.1371/journal.pcbi.1007630)

43. Price MN, Dehal PS, Arkin AP. 2010 Fasttree 2 – approximately maximum-likelihood trees for large alignments. PLoS ONE 5, 1–10. (doi:10.1371/journal.pone.0009490)

44. Bitbol AF. 2018 Inferring interaction partners from protein sequences using mutual information. PLoS Comput. Biol. 14, e1006401. (doi:10.1371/journal.pcbi.1006401)

45. Bialek W. 2012 Biophysics: searching for principles. Princeton, NJ: Princeton University Press.

46. Dietler N, Lugo U, Bitbol AF. 2023 Impact of phylogeny on structural contact inference from protein sequence data. Figshare. (doi:10.6084/m9.figshare.c.6392426)