Supplementary material for "Impact of phylogeny on structural contact inference from protein sequence data"

Nicola Dietler[1,2], Umberto Lupo[1,2], Anne-Florence Bitbol[1,2,*]

1 Institute of Bioengineering, School of Life Sciences, École Polytechnique Fédérale de Lausanne (EPFL), CH-1015 Lausanne, Switzerland
2 SIB Swiss Institute of Bioinformatics, CH-1015 Lausanne, Switzerland
* Corresponding author: anne-florence.bitbol@epfl.ch

Contents

S1 Data generation at equilibrium in the minimal model

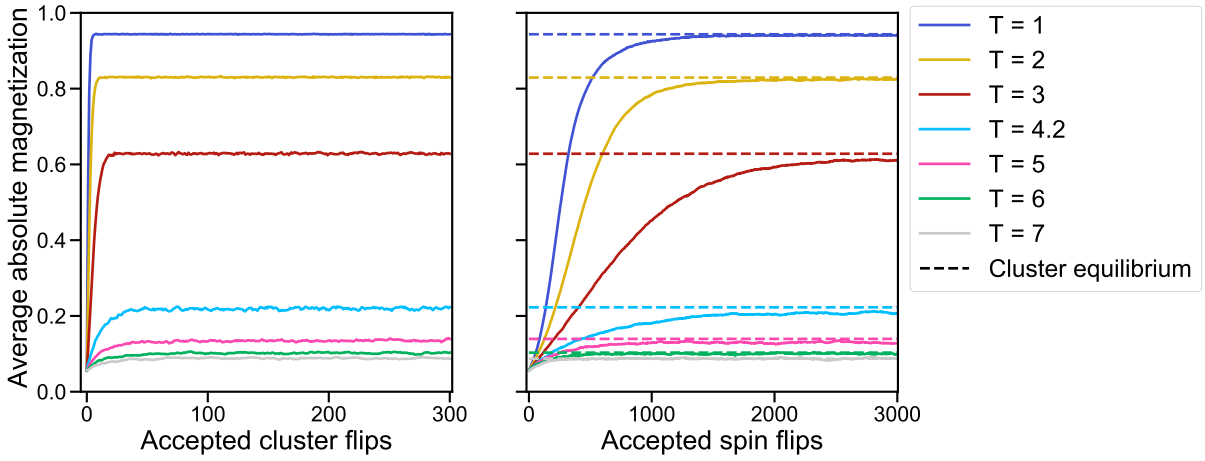

Figure S1: **Generation of independent equilibrium sequences.** The average absolute magnetisation per site is shown versus the number of accepted cluster flips (left panel) or spin flips (right panel) during Metropolis Monte Carlo sampling. Initialisation is made from sequences where each spin is chosen uniformly at random. Then, either a cluster (Wolff) algorithm (left panel) or a single flip algorithm (right panel) is used. Gradually, equilibrium is reached, and magnetisation stabilises on a plateau. This happens faster with the cluster algorithm. The final value reached with this algorithm is shown on the right panel for comparison. The same contact map (Erdős-Rényi graph) as in Figure 2 is used, and different sampling temperatures T are employed. The absolute magnetisation is averaged over all sites in the sequence, and over 2048 sequences, each starting from a different random initialisation.

S2 Impact of small nonzero couplings

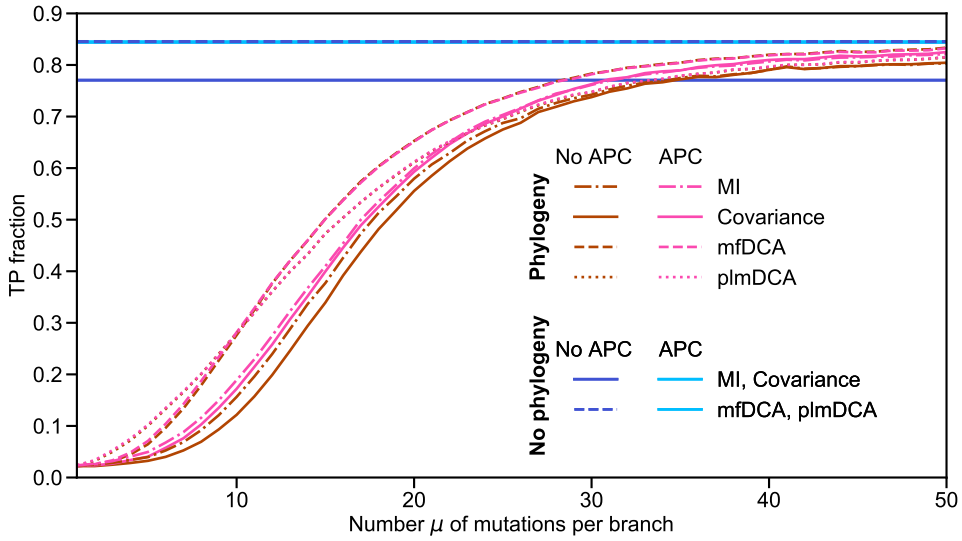

Figure S2: **Impact of phylogeny on contact prediction with small nonzero couplings.** The TP fraction is plotted versus the number μ of mutations per branch of the phylogenetic tree, as in Figure 2. The difference is that here, couplings are distributed according to a Gaussian distribution with mean 0 and standard deviation 1, and then multiplied by -1 if negative, to have only positive couplings as in our minimal model. The number of sites is 200 as usual, but a new Erdős-Rényi graph is generated with 1300 contacts, so that there are 414 couplings greater than or equal to 1, which are considered as the actual contacts to be inferred. This number is very close to the number of contacts $N_{\text{contacts}} = 413$ in the sparse graph of Figure 2. However, there are many small nonzero couplings in addition. All results are averaged over 100 realisations. Note that the performance of local methods with phylogeny and no APC does not tend to that of the corresponding equilibrium case at high μ. This is due to the sampling effect described in section S3.

S3 Impact of conservation on local methods at low T

A surprising feature in the left panel of Figure 3 is that the performance of local methods is worse at equilibrium than for $\mu = 5$ and $\mu = 15$ at low T. Besides, this effect is corrected by the APC (Figure 3, right panel). To understand this, let us examine the way the data is generated. At low T, deep in the ferromagnetic phase, equilibrium sequences mainly comprise aligned spins, with a large majority of either 1 (positive overall magnetisation) or -1 (negative overall magnetisation). On average, half of the independent equilibrium sequences are in each of these classes, leading to very strongly correlated baseline scores for every pair of sites. In contrast, when generating data with phylogeny at low T, all sequences generally have the same overall magnetisation sign, because they all stem from the same equilibrium ancestor, and its magnetisation sign is generally preserved through the phylogeny. The resulting high conservation leads to a baseline of weakly correlated sites, which may allow to detect some signal from the few sites which flipped in the phylogeny, allowing to predict a few contacts. To test this hypothesis, we transform the equilibrium data sets at $T < T_C$ by fully flipping each sequence with negative magnetisation, yielding only sequences with positive magnetisation. Figure S3 shows that inference employing local methods is indeed drastically improved by this transformation, while DCA results are very little impacted, probably partly thanks to the large pseudocount in mfDCA and the regularisation in plmDCA. Indeed, a pseudocount can reduce the value of very high correlations and can also turn a strong conservation into a correlation (see details below), thereby somewhat easing the difficulties encountered for these extreme cases. This data transformation further allows to make a direct link between equilibrium data and data generated with little phylogeny: Figure S3 shows that data generated with $\mu = 50$ behaves similarly as transformed equilibrium data. Importantly, with this transformation, the difference of performance between local and DCA methods for equilibrium data is substantially reduced. In Figure S3, inference using mutual information is almost as good as using DCA for low temperatures. The impact of our data transformation modifying magnetisation signs is in very similar to that of the APC shown in Figure 3. This makes sense, as the APC essentially increases contrast in the score matrices by subtracting the product of means over rows and columns, and can thus partly amend the collective effects observed in equilibrium ferromagnetic sequences. Interestingly, these low-T results show that conservation can help contact inference in some cases.

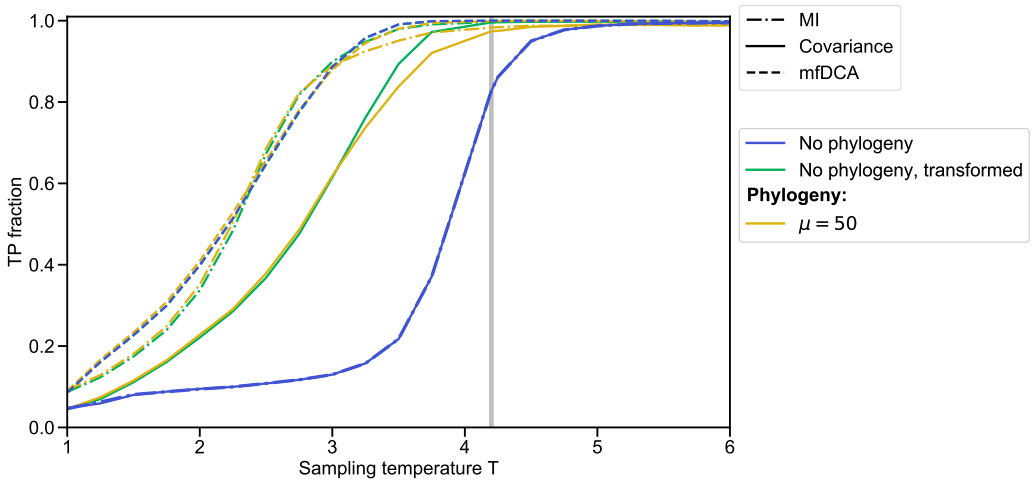

Figure S3: **Impact of data sampling details on contact prediction.** The correctly predicted fraction of contacts (TP fraction) is plotted versus the Monte Carlo sampling temperature T for three different inference methods (Covariance, MI, mfDCA). Note that, for the sake of readability, plmDCA results are not shown, but they behave very similarly as with mfDCA. The equilibrium (no phylogeny) data set employed is the same as in Figure 2. We also consider a transformed data set, built by multiplying by -1 the sequences of this equilibrium data set that have negative magnetisation. The data set for $\mu = 50$ is generated as in Figure 2, using the same contact map (Erdős-Rényi graph), but with $\mu = 50$ (weak phylogeny).

Impact of the pseudocount in simple cases. Considering a sequence with just two sites, the covariance matrix can be written as:

$$C = \begin{pmatrix} C_{11} & C_{12} \\ C_{21} & C_{22} \end{pmatrix}, \tag{1}$$

where $C_{ij} = \langle \sigma_i \sigma_j \rangle - \langle \sigma_i \rangle \langle \sigma_j \rangle$. Adding a pseudocount λ amounts to replacing C_{ij} by:

$$\tilde{C}_{ij} = (1-\lambda)\left[C_{ij} + \lambda \langle \sigma_i \rangle \langle \sigma_j \rangle\right] \text{ for } i \neq j\,, \tag{2}$$

$$\tilde{C}_{ii} = (1-\lambda)^2\left[1 - \langle \sigma_i \rangle^2\right] + \lambda(2-\lambda)\,. \tag{3}$$

In the mean-field approximation, the couplings J can be computed from $\tilde{C}$ via $J = -\tilde{C}^{-1}$. In our two-site case, this yields:

$$\tilde{C}^{-1} = \frac{1}{\det \tilde{C}} \begin{pmatrix} \tilde{C}_{22} & -\tilde{C}_{12} \\ -\tilde{C}_{21} & \tilde{C}_{11} \end{pmatrix}, \tag{4}$$

with

$$\det \tilde{C} = \tilde{C}_{11}\tilde{C}_{22} - \tilde{C}_{12}\tilde{C}_{21}\,. \tag{5}$$

Thus, for instance,

$$J_{12} = \frac{1}{\det \tilde{C}} \tilde{C}_{12} \tag{6}$$

Therefore, in this two-site case, the couplings are equal to the pseudocount-corrected covariance up to the inverse determinant prefactor.

Let us now consider two specific extreme situations in this two-site case. In the first one, $\sigma_1 = \sigma_2 = 1$ in every sequence (perfect conservation), while in the second one, $\sigma_1 = \sigma_2 = 1$ in one half of the sequences and $\sigma_1 = \sigma_2 = -1$ in the other half of the sequences (perfect correlation, but no conservation).

The first situation yields $C_{12}^{(1)} = 0$ and $C_{11}^{(1)} = 0$, leading to a non-invertible covariance matrix without a pseudocount. Including a pseudocount gives $\tilde{C}_{12}^{(1)} = \tilde{C}_{21}^{(1)} = \lambda(1-\lambda)$ (see Eq. 2), which has a maximum at $\lambda = 0.5$, the usual value used in mfDCA. To obtain the couplings, we compute the determinant using Eq. 5 (and Eq. 3, which provides $\tilde{C}_{11}^{(1)} = \tilde{C}_{22}^{(1)} = \lambda(2-\lambda)$). Thus,

$$\det \tilde{C}^{(1)} = \left[\lambda(2-\lambda)\right]^2 - \left[\lambda(1-\lambda)\right]^2 = \lambda^2\,(3-2\lambda)\,. \tag{7}$$

This gives

$$J_{12}^{(1)} = \frac{1}{\det \tilde{C}^{(1)}} \tilde{C}_{12}^{(1)} = \frac{1-\lambda}{\lambda\,(3-2\lambda)}. \tag{8}$$

In the second situation, $C_{12}^{(2)} = 1$ and $C_{11}^{(2)} = 1$, yielding a non-invertible covariance matrix in the absence of a pseudocount. Including a pseudocount gives $\tilde{C}_{12}^{(2)} = 1-\lambda$, while $\tilde{C}_{11}^{(2)} = \tilde{C}_{22}^{(2)} = (1-\lambda)^2 + \lambda(2-\lambda) = 1$. The determinant reads

$$\det \tilde{C}^{(2)} = 1 - (1-\lambda)^2 = \lambda(2-\lambda)\,. \tag{9}$$

This leads to

$$J_{12}^{(2)} = \frac{1}{\det \tilde{C}^{(2)}} \tilde{C}_{12}^{(2)} = \frac{1-\lambda}{\lambda\,(2-\lambda)}. \tag{10}$$

To compare these two situations, let us first consider the ratio of covariances between the two sites

$$\frac{\tilde{C}_{12}^{(1)}}{\tilde{C}_{12}^{(2)}} = \frac{\lambda(1-\lambda)}{1-\lambda} = \lambda\ : \tag{11}$$

we have $0 < \tilde{C}_{12}^{(1)} < \tilde{C}_{12}^{(2)} < 1$ for all values of $\lambda \in]0,1[$. Besides, while in the first situation the two sites have zero covariance in the absence of a pseudocount, we see here that the ratio $\tilde{C}_{12}^{(1)}/\tilde{C}_{12}^{(2)}$ is equal to $1/2$ for the usual choice $\lambda = 0.5$ (and even becomes 1 for $\lambda \to 1$). Let us now turn to the couplings between these two sites. Their ratio is

$$\frac{J_{12}^{(1)}}{J_{12}^{(2)}} = \frac{1-\lambda}{\lambda\,(3-2\lambda)} \frac{\lambda\,(2-\lambda)}{1-\lambda} = \frac{2-\lambda}{3-2\lambda}\ : \tag{12}$$

we have $0 < J_{12}^{(1)} < J_{12}^{(2)}$ for all values of $\lambda \in]0,1[$. More precisely, the ratio $J_{12}^{(1)}/J_{12}^{(2)}$ increases with λ, being equal to 2/3 for $\lambda \to 0$ and to 1 for $\lambda \to 1$, and to 3/4 for the usual choice $\lambda = 0.5$.

Therefore, using a pseudocount transforms the fully-conserved case from no covariance and an indetermination for the coupling to a covariance value and a coupling value that are finite and comparable to the fully-correlated but not-conserved case. The magnitude of these effective covariances and couplings coming from conservation is larger if the value of λ is increased. In particular, for the usual value $\lambda = 0.5$, the effective coupling represents 3/4 of those coming from conservation-free correlation.

S4 Comparison with Ref. [1]

In Ref. [1], the performance of mfDCA and covariance was compared on equilibrium data sets generated in a minimal model similar to ours, at different sampling temperatures. However, in Ref. [1], covariance was found to perform better than mfDCA for all values of T in data sets of similar size as ours. One difference is that a substantially denser graph was employed. Thus, to make a more in-depth comparison with Ref. [1], we generated data using an Erdős-Rényi graph with $q = 0.2$ (recall that $q = 0.02$ in the rest of our work, q being the probability that two nodes are connected in the graph). Moreover, we started by using no pseudocounts and assessing contact prediction performance via the AUC (area under the receiver operating characteristic), as in Ref. [1]. Consistently with Ref. [1], we find that covariance and mutual information then perform better than DCA for equilibrium data (see top left panels of Figures S4 and S5). The same result holds when using TP fraction to assess performance (see top middle panels of Figures S4 and S5). Note that TP fraction is most often used in the DCA field, which is why we employ it here, but that the AUC has the advantage of not depending on a threshold (here, a number of predicted contacts). While these results are obtained without regularisation, as in Ref. [1], we observe that high pseudoucount or regularisation strength values allow DCA to reach better prediction performance and outperform covariance and mutual information for low temperatures, while performing similarly for larger temperatures. Interestingly, even higher regularisation strengths are required for this denser graph (see Figures S4 and S5). This illustrates the strong importance of regularisation for the performance of DCA. A more detailed analysis of the impact of pseudocount or regularisation strength on the inference of contacts by DCA is shown in Figure S6. Hence, our results are consistent with those of Ref. [1], and the apparent differences arise mainly from regularisation. Furthermore, Figures S4 and S5 confirm that global methods (DCA) handle phylogeny better than local ones (covariance and mutual information) for this denser graph, as for our sparse graph (see Fig. 3). We also studied the impact of APC on inference in the denser graph (see Figures S4 and S5), and found that it slightly deteriorates the inference performance of global methods, especially on data containing phylogeny, while it improves that of local methods, especially on equilibrium data, consistently with our results above.

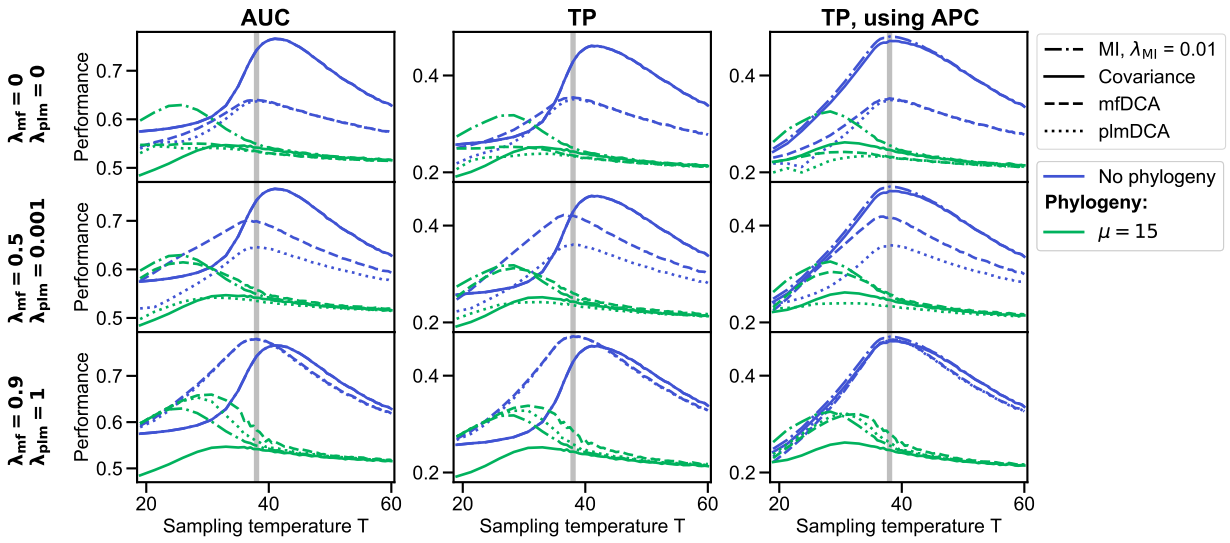

Figure S4: **Contact prediction performance for a denser random graph on equilibrium and phylogenetic data.** The results of different performance assessments, namely the area under the receiver operating characteristic curve (AUC), the TP fraction and the TP fraction using APC for four inference methods (C, MI, mfDCA, plmDCA) is shown on two different data sets. Data has been generated at equilibrium (no phylogeny) and with phylogeny (here $\mu = 15$) in the same way as in Figure 3 but using a denser contact map (Erdős-Rényi graph with probability $q = 0.2$ instead of $q = 0.02$). Each line of plots in this figure represent the inference with different values of pseudocounts and regularisation parameters used in the DCA methods, the actual values are shown on the left in bold. The value of the pseudocount is fixed for MI and none is used for covariance. The top left panel recovers the result found in [1].

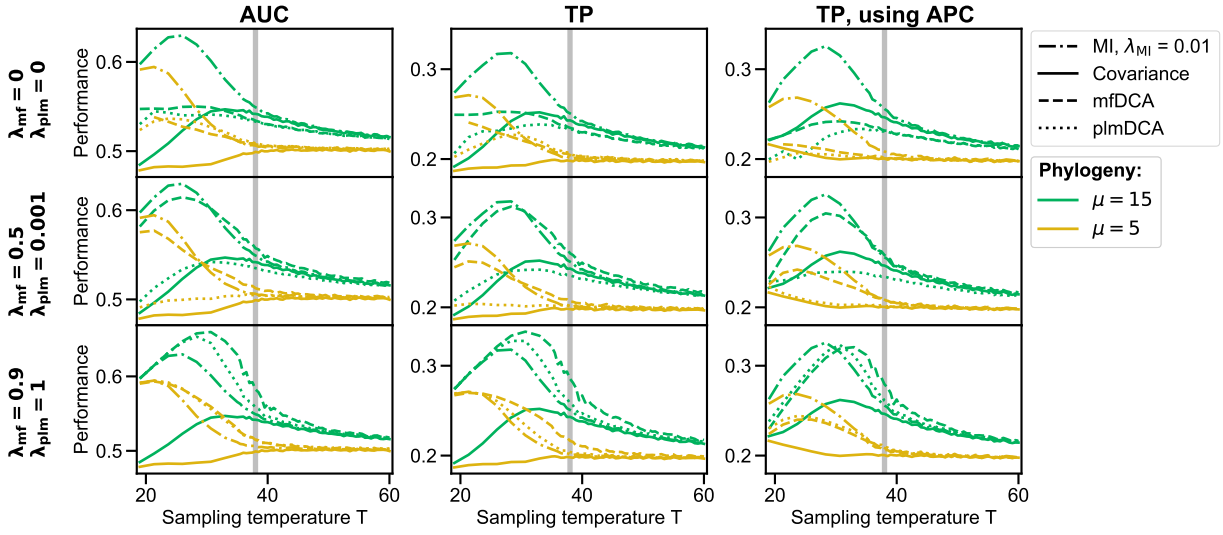

Figure S5: **Contact prediction performance for a denser random graph on two different phylogenetic data sets.** Same plots as in Figure S4, but for two different phylogenetic data sets, one with $\mu = 5$ and the other one with $\mu = 15$.

S5 Impact of pseudocount and regularisation on contact prediction

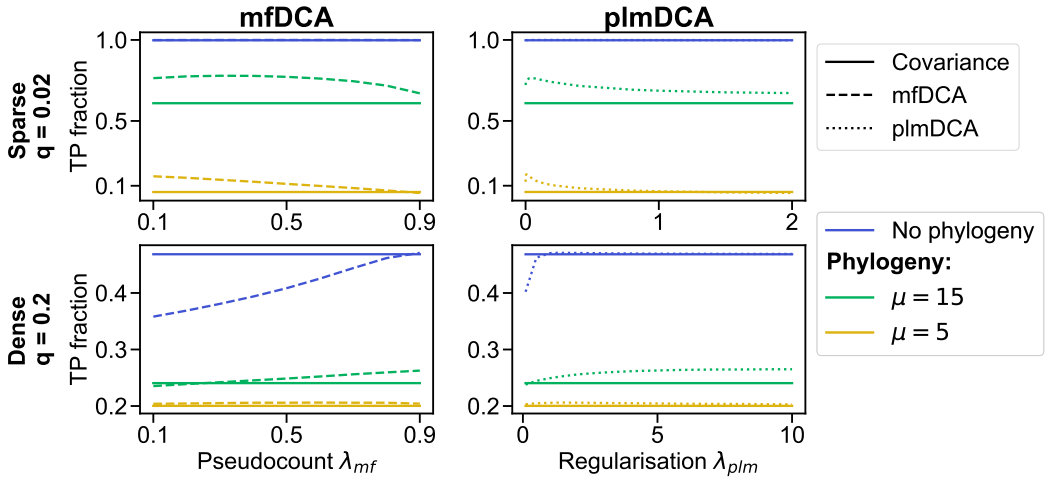

Figure S6: **Impact of pseudocount and regularisation on contact prediction in sparse and denser graphs.** The fraction of correctly predicted contacts (TP fraction) is plotted versus the pseudocount for mfDCA, or the regularisation strength for plmDCA, in Erdős-Rényi graphs with two different probabilities of contact ($q = 0.02$ and $q = 0.2$). In each case, the performance of contact prediction by the covariance matrix is also shown for reference. Three data sets are presented in each case: data generated independently at equilibrium (no phylogeny), and data generated with phylogeny at $\mu = 15$ and $\mu = 5$. For the sparse graph, data is generated as in Figure 2 at $T = 5$. For the denser graph, data is generated as in Figure S4 at $T = 40$. All results are averaged over 100 realisations.

S6 Impact of graph properties on coevolution scores

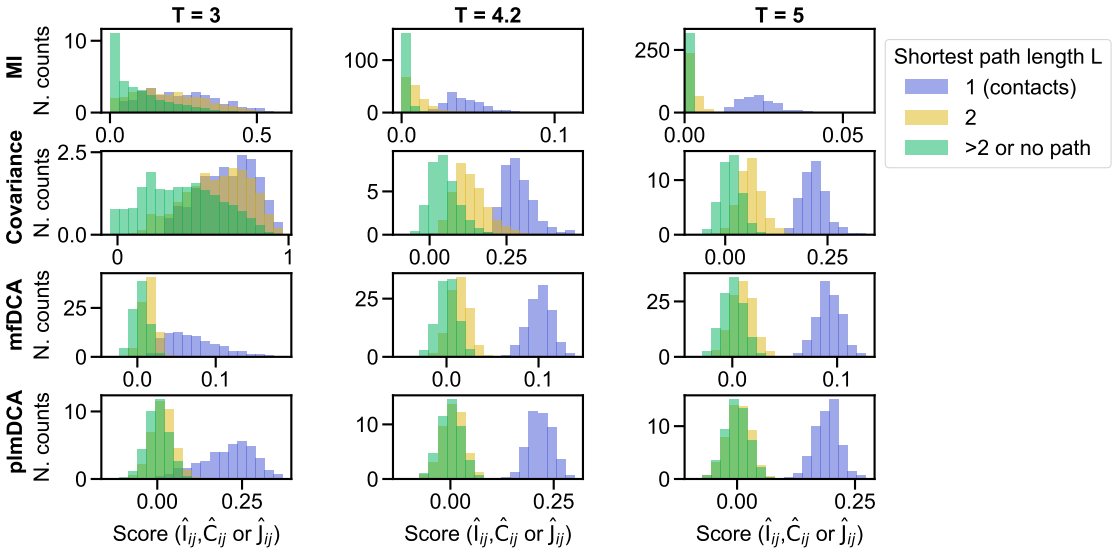

Figure S7: **Impact of graph properties on coevolution scores for equilibrium sequences.** Histograms of coevolution scores for all pairs of sites (i, j) are shown for equilibrium data sampled at three different temperatures: $T = 3 < T_C$ (left), $T = 4.2 \simeq T_C$ (center), and $T = 5 > T_C$ (right), for four inference methods (MI, Covariance, mfDCA, plmDCA) without APC. Pairs of sites are split into three categories according to the length L of the shortest path connecting them in the graph: contacts, $L = 1$; first indirect neighbours, $L = 2$; more distant ($L > 2$ or isolated sites). Normalised counts (N. counts) are shown – note that there are many more non-contact than contact pairs. Data sets of 2048 sequences each are generated at equilibrium using the cluster algorithm (see Figure S1), using the same contact map (Erdős-Rényi graph) as in Figure 2.

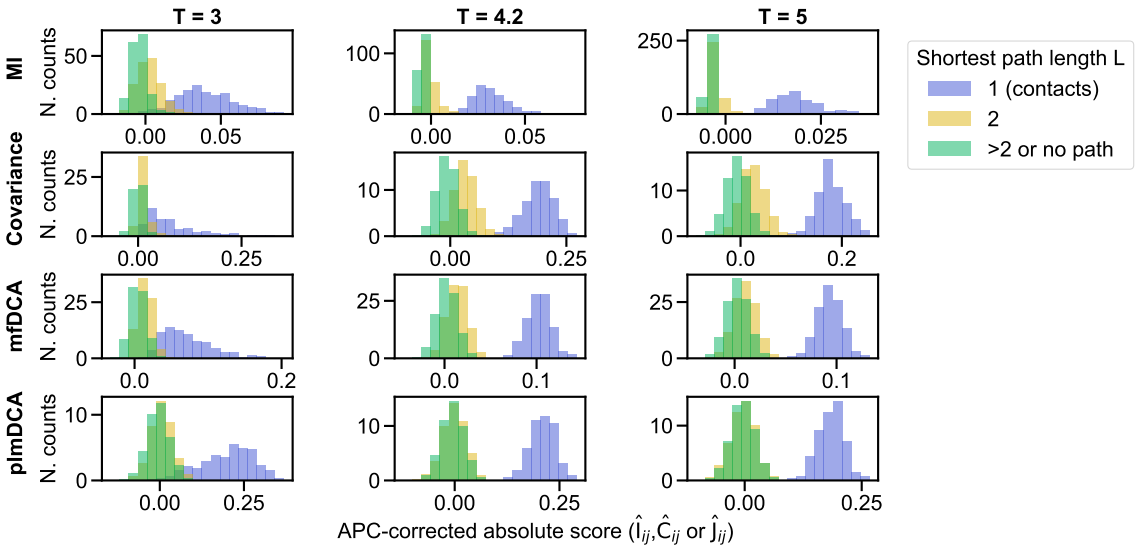

Figure S8: **Impact of graph properties on APC-corrected absolute scores for equilibrium sequences.** Same as in Figure S7 except that the APC-corrected absolute values of the scores are reported instead of the raw scores.

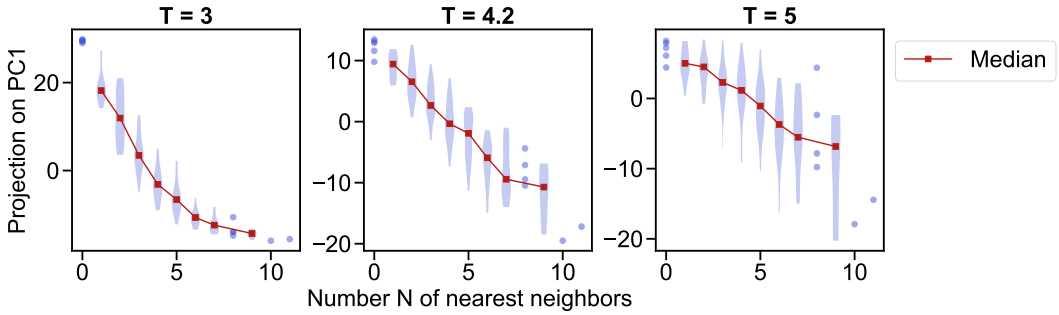

Figure S9: **Impact of graph properties on data covariance for equilibrium sequences.** Violin plots of the projection of each site on the first principal component are shown versus the number N of nearest neighbours of these sites. In each panel, principal component analysis (PCA) is performed on a data set of equilibrium sequences (where sequences are features and sites are observations – in other words, the matrix of covariances between sequences, which is an $M \times M$ size matrix, where M is the number of sequences, is diagonalised). The first principal component is the direction of largest variance of the sites (top "eigensite"). Three data sets of 2048 sequences each are sampled at equilibrium at three temperatures: $T = 3 < T_C$ (left), $T = 4.2 \simeq T_C$ (center), and $T = 5 > T_C$ (right), employing the cluster algorithm (see Figure S1), and using the same contact map (Erdős-Rényi graph) as in Figure 2. Ensembles of 5 data points or fewer are represented with individual markers instead of Gaussian kernel-smoothed histograms.

S7 Impact of phylogeny on coevolution scores and on conservation

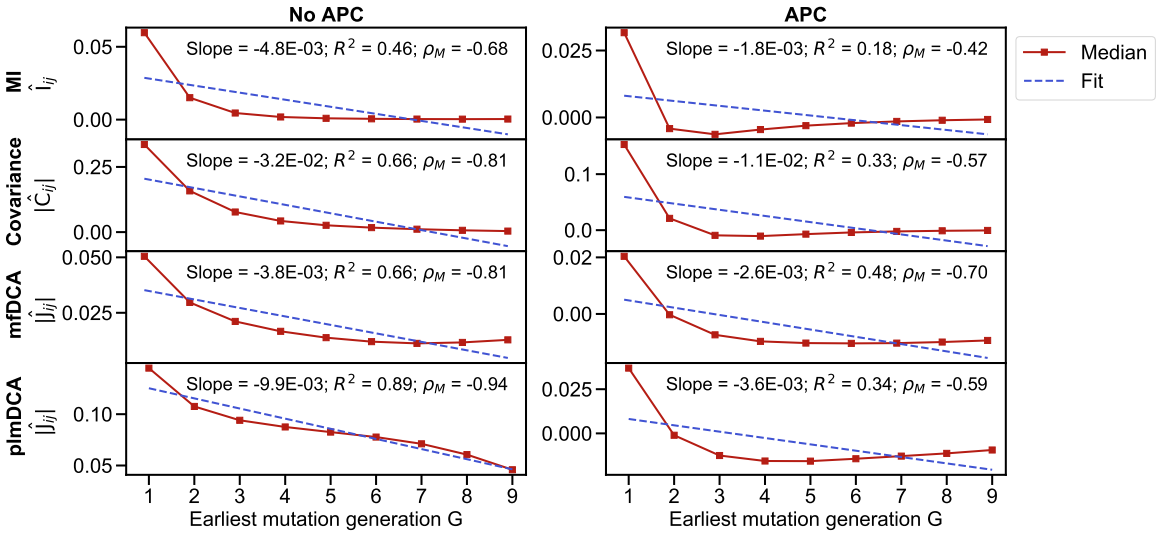

Figure S10: **Impact of phylogeny on the median of the coevolution scores.** The median of the violin plots in Figure 5 is plotted versus the earliest generation G without (left panel) and with (right panel) the APC correction on the scores. A linear fit of the median is performed in each case, and the slope of the fit, the coefficient of determination R^2 and the Pearson correlation ρ_M between the median score and G are shown. The data is the same as in Figure 5.

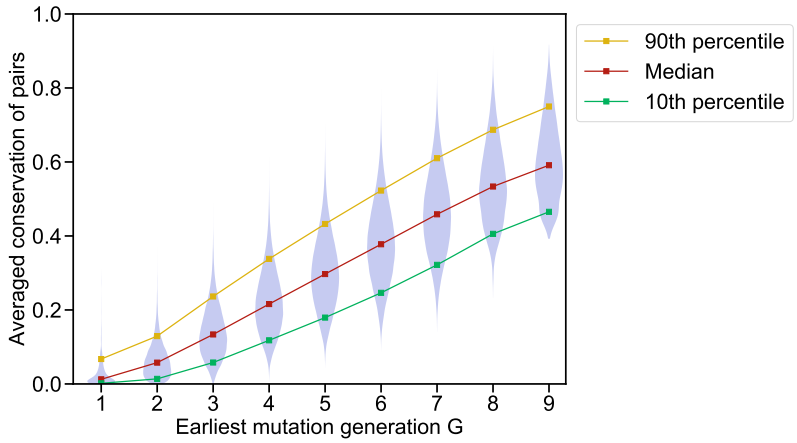

Figure S11: **Impact of phylogeny on conservation scores.** Violin plots of the average conservation scores of all pairs of sites (i, j) are shown versus the earliest generation G where both i and j have mutated with respect to the ancestral sequence. The conservation of i is defined as the difference of the maximum possible entropy of a site (1 bit) and the estimated entropy of i (in bits), computed using frequencies instead of probabilities. The average conservation of a pair (i, j) is the mean of the scores of i and j. Data is generated as in Figure 2, using the same contact map (Erdős-Rényi graph), and parameters $T = 5$ and $\mu = 5$. Couplings are inferred on 100 data sets comprising 2048 sequences each, and aggregated. The squared Pearson correlation between average conservation and earliest mutation generation is $r^2 = 0.53$.

S8 Natural and more realistic data

		MSA			PDB structure		Contact density	
Pfam ID	Family name	ℓ	M	$M_{\text{eff}}^{(PR)}$	ID	Resol.	4 Å	8 Å
PF00072	Response_reg	112	73063	15090	3ILH	2.59 Å	0.02	0.13
PF00512	HisKA	66	154998	8980	3DGE	2.80 Å	0.01	0.13
PF00595	PDZ	82	71303	1419	1BE9	1.82 Å	0.04	0.18
PF02518	HATPase_c	111	80714	16058	3G7E	2.20 Å	0.02	0.11

Table S1: **Pfam families and MSAs considered in this work.** For each family, we considered the Pfam full alignments ("MSA"). For each of these MSAs, we report the length ℓ (number of amino acid sites), depth M (number of sequences) and the effective depth $M_{\text{eff}}^{(PR)}$ from the phylogenetic reweighting done in the `PlmDCA` package (https://github.com/pagnani/PlmDCA). The PDB structures used as reference structures for each Pfam family, and their resolutions, are reported. The density of the experimental contact maps (i.e. fraction of residue pairs i, j that are in contact) is also shown, with two all-atom Euclidean distance cutoffs at 4 Å and 8 Å, excluding residue pairs at positions i, j with $|i - j| \leq 4$.

	MI				plmDCA			
Pfam ID	Raw	PR	APC	PR+APC	Raw	PR	APC	PR+APC
PF00072	0.36	0.52	0.59	0.70	0.62	0.74	0.77	0.81
PF00512	0.21	0.27	0.33	0.30	0.33	0.38	0.42	0.43
PF00595	0.22	0.46	0.53	0.66	0.27	0.39	0.52	0.70
PF02518	0.22	0.25	0.39	0.41	0.34	0.42	0.48	0.48

Table S2: **Impact of phylogenetic corrections on contact prediction in natural MSAs.** The four Pfam protein families introduced in Table S1 were used. Contacts are predicted using mutual information (local) and plmDCA (global), without or with two different phylogenetic corrections. Performance of contact prediction is assessed as the fraction of true positive contacts among $N_{\text{pred}} = 2\ell$ predicted contacts (see Table S3), which coincides with the PPV. "Raw" means that no phylogenetic correction is used; "PR" indicates that the phylogenetic reweighting correction of the frequencies is used; "APC" means that the APC correction is used; "PR+APC" means that both phylogenetic reweighting and APC are used, which is standard in DCA.

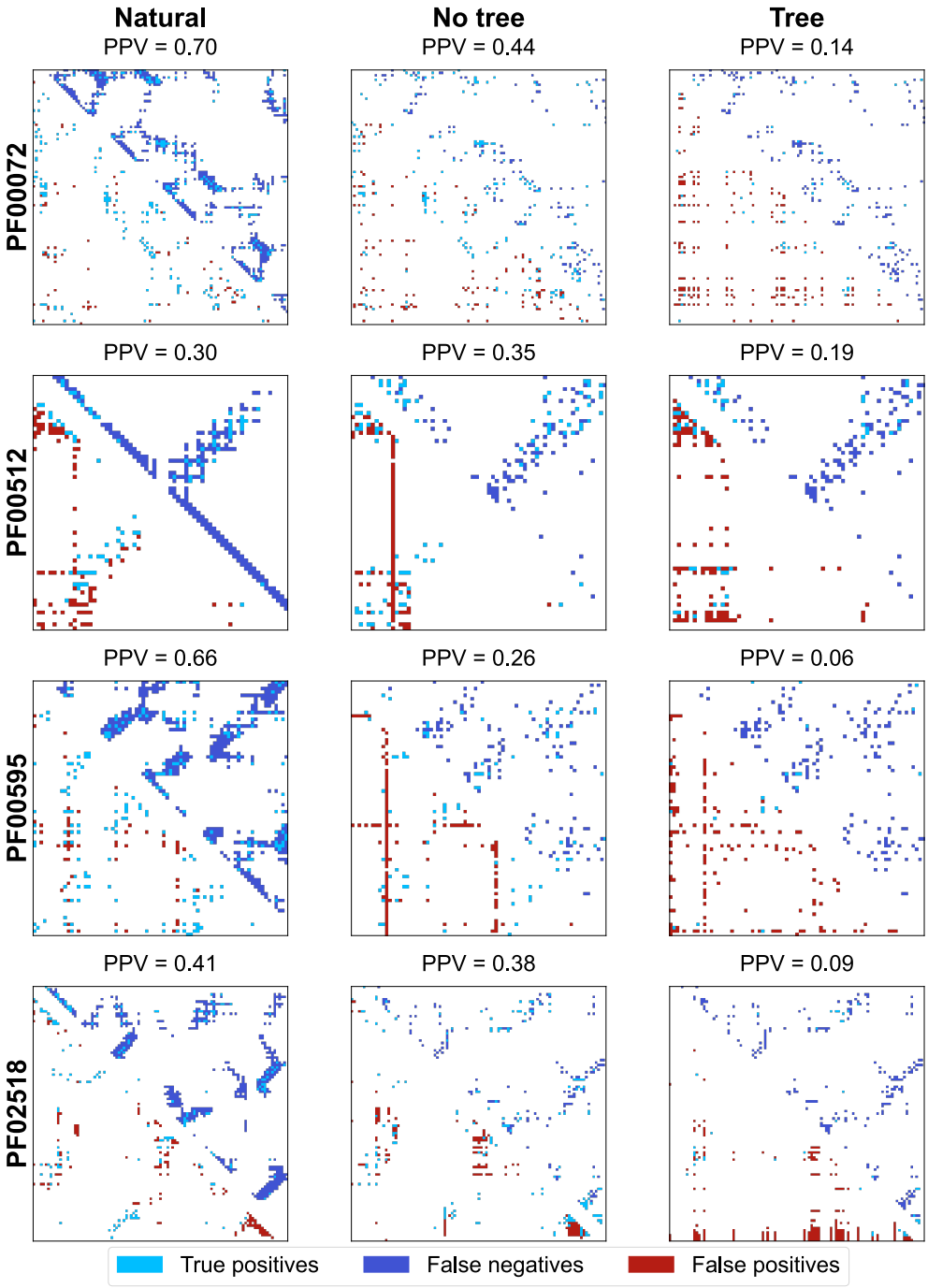

Figure S12: **Contact maps predicted by mutual information from natural and realistic MSAs.** Each panel shows the ground truth in the upper triangular part and the mutual information (MI)-based inference in the lower triangular part. Inference is performed using the phylogenetic reweighting (PR) and APC corrections. For each Pfam family, the left column shows an experimental contact map as ground truth and the MI-based inference performed on the natural MSA. Experimental contact maps use the PDB structures listed in table S1, with an all-atom Euclidean distance cutoff of 8 Å, excluding residue pairs at positions i, j with $|i - j| \leq 4$. The middle (resp. right) column presents the contact maps inferred from the natural sequences by bmDCA as ground truth and the MI-based inference performed on synthetic sequences generated at equilibrium (resp. along a phylogenetic tree inferred from the natural MSA) using the bmDCA model inferred from the natural MSA (see Methods, "Generating more realistic sequences"). In all cases, the top 2ℓ scores are predicted as contacts, and the associated PPV (or true positive fraction) is shown.

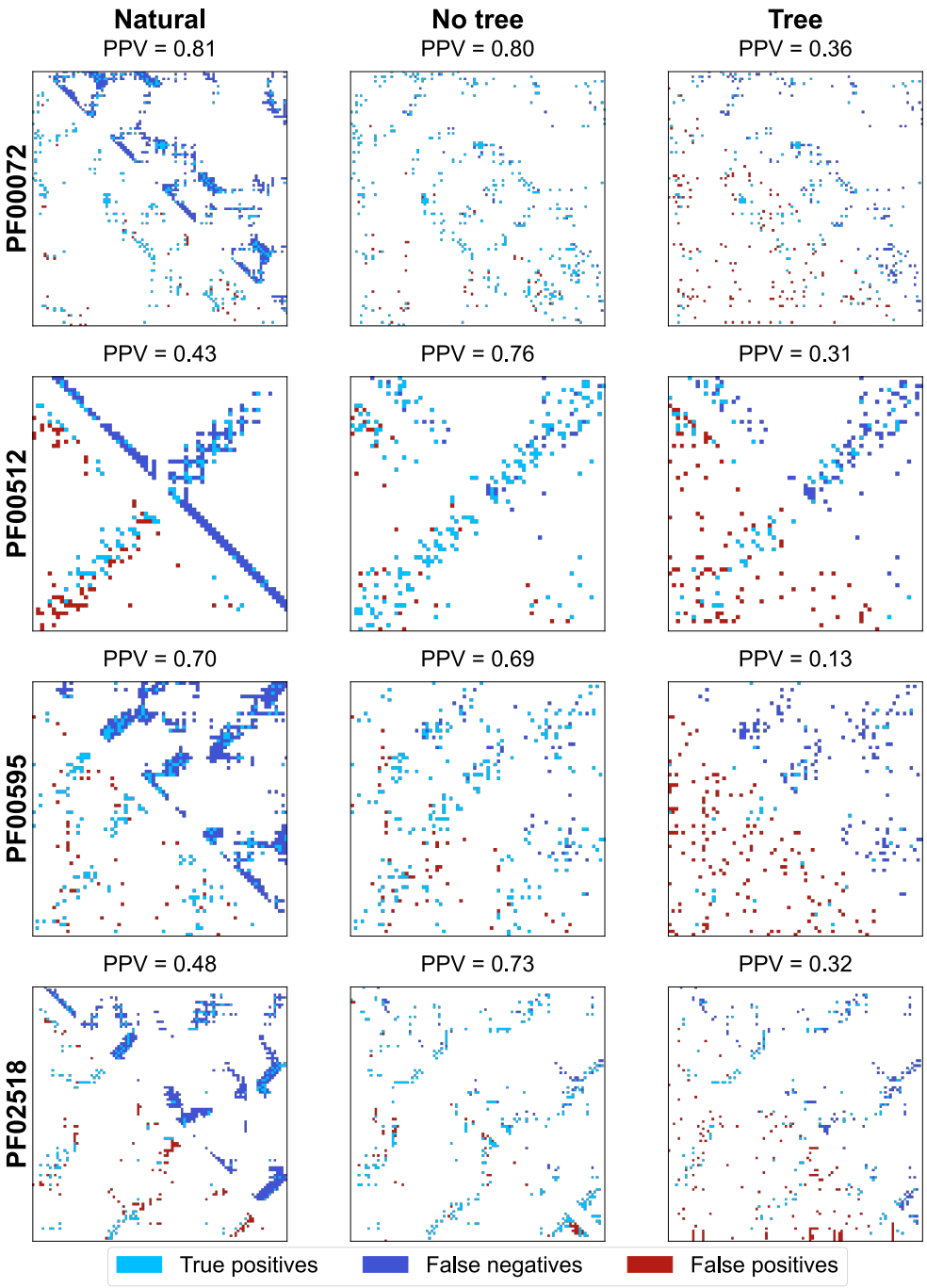

Figure S13: **Contact maps predicted by plmDCA from natural and more realistic MSAs.** Same as in figure S12, but using plmDCA instead of MI to infer contacts. Here too, phylogenetic reweightings are used, as well as the APC correction.

	Method			Natural		Equilibrium		Tree	
Pfam ID	Score	Correction	N_{pred}	PPV	$\text{FP}_{L=2}$	PPV	$\text{FP}_{L=2}$	PPV	$\text{FP}_{L=2}$
PF00072	MI	Raw	224	0.36	89	0.23	48	0.13	32
	MI	PR+APC	224	0.70	60	0.44	34	0.14	30
	plmDCA	Raw	224	0.62	55	0.65	33	0.26	21
	plmDCA	PR+APC	224	0.81	30	0.80	19	0.36	22
PF00512	MI	Raw	132	0.21	49	0.32	36	0.14	24
	MI	PR+APC	132	0.30	50	0.35	20	0.19	40
	plmDCA	Raw	132	0.33	47	0.55	29	0.18	29
	plmDCA	PR+APC	132	0.43	46	0.76	12	0.31	27
PF00595	MI	Raw	164	0.22	58	0.41	54	0.06	20
	MI	PR+APC	164	0.66	51	0.26	14	0.06	15
	plmDCA	Raw	164	0.27	55	0.62	34	0.07	33
	plmDCA	PR+APC	164	0.70	38	0.69	17	0.13	20
PF02518	MI	Raw	222	0.22	73	0.22	36	0.12	19
	MI	PR+APC	222	0.41	66	0.38	47	0.09	11
	plmDCA	Raw	222	0.34	71	0.58	32	0.27	26
	plmDCA	PR+APC	222	0.48	61	0.73	18	0.32	13

Table S3: **Contact prediction performance and indirect correlations in natural and realistic data sets.** Performance of contact prediction is assessed as the fraction of true positive contacts among $N_{\text{pred}} = 2\ell$ predicted contacts, which coincides with the PPV. Mutual information (local) and plmDCA (global) are used to predict contacts, either with no phylogenetic correction ("Raw") or with both phylogenetic reweighting and APC ("PR+APC"). For each of the four Pfam families on Table S1, three different data sets are compared: natural MSAs, equilibrium bmDCA-generated MSAs and MSAs generated using bmDCA along an inferred phylogenetic tree (see Methods, "Generating more realistic sequences"). In addition to the PPVs, the number of false positive pairs with a shortest path length of $L = 2$ (i.e., indirect correlations of order 1) are shown.

References

[1] V. Ngampruetikorn, V. Sachdeva, J. Torrence, J. Humplik, D. J. Schwab, and S. E. Palmer. Inferring couplings in networks across order-disorder phase transitions. *Phys. Rev. Research*, 4:023240, Jun 2022.