Improved contact prediction in proteins: Using pseudo-likelihoods to infer Potts models

Magnus Ekeberg[1], Cecilia Lövkvist[3],Yueheng Lan[4], Martin Weigt[5], Erik Aurell[2,3,†*]

(Dated: October 17, 2012)

Spatially proximate amino acid in a protein tend to co-evolve. A protein's 3D-structure hence leaves an echo of correlations in the evolutionary record. Reverse engineering 3D-structures from such correlations is an open problem in structural biology, pursued with increasing vigor as more and more protein sequences continue to fill the data banks. Within this task lies a statistical inference problem, rooted in the following: correlation between two sites in a protein sequence can arise from firsthand interaction, but can also be network-propagated via intermediate sites; observed correlation is not enough to guarantee proximity. To separate direct from indirect interactions is an instance of the general problem of *inverse statistical mechanics*, where the task is to learn model parameters (fields, couplings) from observables (magnetizations, correlations, samples), in large systems. In the context of protein sequences, the approach is referred to as *direct-coupling analysis* (Weigt *et al*, 2009). Here we show that the pseudo-likelihood method, applied to 21-state Potts models describing the statistical properties of families of evolutionarily related proteins significantly outperforms existing approaches to the direct-coupling analysis, the latter being based on standard mean-field techniques. The results are verified using known crystal structures of specific sequence instances of various protein families.

PACS numbers: 02.50.Tt – Inference methods, 87.10.Vg – Biological information, 87.15.Qt – Sequence analysis, 87.14.E- – Proteins

I. INTRODUCTION

In biology, new and refined experimental techniques have triggered a rapid increase in data availability during the last few years. Such progress needs to be accompanied by the development of appropriate statistical tools to treat growing data sets. An example of a branch undergoing intense growth in the amount of existing data is *protein structure prediction* (PSP), which, due to the strong relation between a protein's structure and its function, is one central topic in biology. As we shall see, one can accurately estimate the 3D-structure of a protein by identifying which amino-acid positions in its chain are statistically coupled over evolutionary time scales [1–5]. Much of the experimental output is today readily accessible through public data bases such as Pfam [6], which collects over 13,000 families of evolutionarily related protein domains, the largest of them containing more than 2×10^5 different amino-acid sequences. Such databases allow researchers to easily access data, to extract information from it and to confront their results.

A recurring difficulty when dealing with interacting systems is distinguishing direct interactions from interactions mediated via multi-step paths across other elements. Correlations are in general straightforward to compute from raw data, whereas parameters describing the true causal ties are not. The network of direct interactions can be thought of as hidden beneath observable correlations, and untwisting it is a task of inherent intricacy. In PSP, using mathematical means to dispose of the network-mediated correlations is necessary to get optimal results [1, 2, 7–9] and yields improvements worth the computational strain put on the analysis. This approach to PSP, which we will refer to as *direct-coupling analysis* (DCA), is the focus of this paper.

In a more general setting, the problem of inferring interactions from observations of instances amounts to *inverse statistical mechanics*, a field which has been intensively pursued in statistical physics over the last decade [10–23]. Similar tasks were earlier formulated in Statistics and Machine Learning where they have been called *model learning* and *inference* [24–27]. To illustrate this concretely, let us start from an Ising model

$$P(\sigma_1, \ldots, \sigma_N) = \frac{1}{Z} \exp\left(\sum_{i=1}^{N} h_i \sigma_i + \sum_{1 \leq i < j \leq N} J_{ij} \sigma_i \sigma_j \right) \tag{1}$$

and its magnetizations $m_i = \partial_{h_i} \log Z$ and connected correlations $c_{ij} = \partial_{J_{ij}} \log Z - m_i m_j$. Counting the number of observables (m_i and c_{ij}) and the number of parameters (h_i and J_{ij}) it is clear that perfect knowledge of the magnetizations and correlations should suffice to determine the external fields and the couplings exactly. It is, however, also clear that such a process must be computationally expensive, since it requires the computation of the partition function Z for an arbitrary set of parameters. The exact but iterative procedure known as Boltzmann machines [28] does in fact work on small systems,

*[1], Engineering Physics Program, KTH Royal Institute of Technology, 100 77 Stockholm, Sweden, [2] ACCESS Linnaeus Center, KTH, Sweden, [3] Department of Computational Biology, AlbaNova University Center, 106 91 Stockholm, Sweden, [4] Department of Physics, Tsinghua University, Beijing 100084, P. R. China, [5] Université Pierre et Marie Curie, UMR7238 – Laboratoire de Génomique des Microorganismes, 15 rue de l'Ecole de Médecine, 75006 Paris, France,[†]eaurell@kth.se

but it is out of question for the problem sizes considered in this paper. On the other hand the naive mean-field equations of (1) read [29–31]:

$$\tanh^{-1} m_i = h_i + \sum_j J_{ij} m_j \qquad (2)$$

From (2) and the fluctuation-dissipation relations an equation can be derived connecting the coupling coefficients J_{ij} and the correlation matrix $\mathbf{c} = (c_{ij})$ [10]:

$$J_{ij} = -(\mathbf{c}^{-1})_{ij} \qquad (3)$$

Equations (2) and (3) exemplify typical aspects of inverse statistical mechanics, and inference in large systems in general. On one hand, the parameter reconstruction using these two equations is *not exact*. It is only *approximate*, because the mean-field equations (2) are themselves only approximate. It also demands reasonably good sampling, as the matrix of correlations is not invertible unless it is of full rank, and small noise on its $\mathcal{O}(N^2)$ entries may result in large errors in estimating the J_{ij}. On the other hand, this method is *fast*, as fast as inverting a matrix, because one does not need to compute Z. Except for mean-field methods as in (2), approximate methods recently used to solve the inverse Ising problem can be grouped as *expansion in correlations and clusters* [16, 19], methods based on the *Bethe approximation* [17, 18, 20–22], and the *pseudo-likelihood method* [23, 27].

For PSP it is not the Ising model but a 21-state Potts model which is pertinent [1]. But which of all the inference methods in inverse statistical mechanics, machine learning or statistics is most suitable for treating real protein sequence data? In how far do the test results obtained for independently generated equilibrium configurations of Potts models translate to the case of protein sequences, which are neither independent nor equilibrium configurations of any well-defined statistical-physics model? The main goal of the this paper is to move towards an answer to this question by showing that the pseudo-likelihood method is a very powerful means to perform DCA, with a prediction accuracy considerably out-performing methods previously assessed.

The paper is structured as follows: in Section II we review the ideas underlying PSP and DCA and explain the biological hypotheses linking protein 3D-structure to correlations in amino-acid sequences. We also review earlier approaches to DCA. In Section III we describe the Potts model in the context of DCA and the properties of exponential families. We further detail a maximum likelihood (ML) approach as brought to bear on the inverse Potts problem and discuss in more detail why this is impractical for realistic system sizes, and we introduce, similarly to (3) above, the inverse Potts mean-field model algorithm for the DCA (mfDCA) and a pseudo-likelihood maximization procedure (plmDCA). This section also covers algorithmic details of both models such as regularization, sequence re-weighting and the choice of interaction scores. In Section IV, we present results from prediction experiments using mfDCA and plmDCA assessed against known crystal structures. In Section V we summarize our findings, put their implications into context, and discuss possible future developments. Appendices contain additional material supporting the main text.

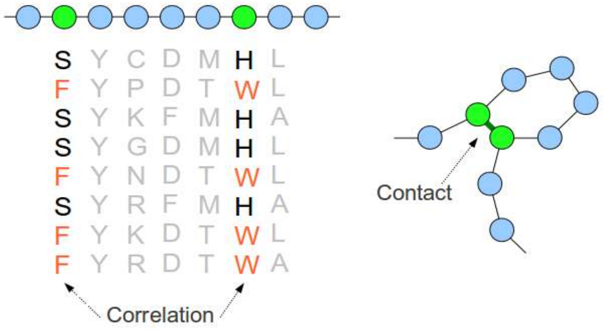
II. PROTEIN STRUCTURE PREDICTION AND DIRECT-COUPLING ANALYSIS

Proteins are essential players in almost all biological processes. Primarily proteins are linear chains, each site being occupied by one out of 20 different amino acids. Their function relies, however, on the protein *fold*, which refers to the 3D conformation into which the amino-acid chain curls. This fold guarantees, e.g., that the right amino-acids are exposed in the right positions at the protein surface, thus forming biochemically active sites, or that the correct pairs of amino acids are brought into contact to keep the fold thermodynamically stable.

Experimentally determining the fold, using for example X-ray crystallography or NMR tomography, is still today a rather costly and time-consuming process. On the other hand, every newly sequenced genome results in a large number of newly predicted proteins. The number of sequenced organisms has by now exceeded 3,700, and continues to grow exponentially (`genomesonline.org` [32]). The most prominent database for protein sequences, Uniprot (`uniprot.org` [33]), lists about 25,000,000 different protein sequences, whereas the number of experimentally determined protein structures is only around 85,000 (`pdb.org` [34]).

However, the situation of structural biology is not as hopeless as these numbers might suggest. First, proteins have a modular architecture, they can be subdivided into *domains* which, to a certain extent, fold and evolve as a unit. Second, domains of a common evolutionary origin, i.e. so-called *homologous* domains, are expected to almost share their 3D structure and to have related function. They can therefore be collected in *protein domain families*: The Pfam data base (`pfam.sanger.ac.uk` [6]) lists almost 14,000 different domain families, the number of the sequences collected in each family ranges roughly from $10^2 - 10^5$. In particular the larger families with more than 1,000 members are of interest to us, as we will argue that their natural sequence variability contains important statistical information about the 3D structure of its member proteins, and can be exploited to successfully address the PSP problem.

Two types of data accessible via the Pfam database are especially important to us. The first is the *multiple sequence alignment* (MSA), a table of the amino acid sequences of all the protein domains in the family lined up to be as similar as possible. A (small and illustrative) example is shown in Fig. 1 (left panel). Normally, not all members of a family can be lined up perfectly, and



S Y C D M H L
F Y P D T W L
S Y K F M H A
S Y G D M H L
F Y N D T W L
S Y R F M H A
F Y K D T W L
F Y R D T W A

FIG. 1: Left panel: small MSA with two positions of correlated amino-acid occupancy. Right panel: hypothetical corresponding spatial conformation, bringing the two correlated positions into direct contact.

the alignment therefore contains both amino acids and *gaps*. At some positions, an alignment will be highly specific (cf. the second, fully conserved column in Fig. 1), while on others it will be more variable. The second data type concerns the *crystal structure* of one or several members of a family. Not all families provide this second type of data. We will discuss its use for an *a posteriori* assessment of our inference results in detail in Sec. IV.

The Potts-model based inference uses only the first data type, i.e. sequence data. Small spatial separation between amino acids in a protein, cf. the right panel of Fig. 1, encourages co-occurrence of favorable amino-acid combinations, cf. the left panel of Fig. 1. This spices the sequence record with correlations, which can be reliably determined in sufficiently large MSAs. However, the use of such correlations for predicting 3D contacts as a first step to solve the PSP problem remained of limited success [35–37], since they can be induced both by direct interactions (amino acid A is close to amino acid B), and also by indirect interactions (amino acids A and B are both close to amino acid C). Lapedes *et al.* [38] were the first to address, in a purely theoretical setting, these ambiguities of a correlation-based route to protein sequence analysis, and these authors also outline a maximum-entropy approach to get at direct interactions. Weigt *et al.* [1] successfully executed this program subsequently called *direct-coupling analysis*: the accuracy in predicting contacts strongly increases when direct interactions are used instead of raw correlations.

To computationally solve the task of inferring interactions in a Potts model, [1] employed a generalization of the iterative message-passing algorithm *susceptibility propagation* previously developed for the inverse Ising problem [17]. Methods in this class are expected to outperform mean-field based reconstruction methods similar to (3) if the underlying graph of direct interactions is locally close to tree-like, an assumption which may or may not be true in a given application such as PSP. A substantial draw-back of susceptibility propagation as used in [1] is that it requires a rather large amount of auxiliary variables, and that DCA could therefore only be carried out on not too long protein sequences. In [2] this obstacle was overcome by using instead a simpler mean-field method, *i.e.* the generalization of (3) to a 21-state Potts model. As discussed in [2], this broadens the reach of the DCA to practically all families currently in Pfam, it improves the computational speed by a factor of about 10^3–10^4, and it appears also more accurate than the susceptibility-propagation based method of [1] in predicting contact pairs. The reason behind this third advantage of mean-field over susceptibility propagation as an approximate method of DCA is unknown at this time.

Pseudo-likelihood maximization (PLM) is an alternative method developed in mathematical statistics to approximate maximum likelihood inference, which breaks down the *a priori* exponential time-complexity of computing partition functions in exponential families [39]. On the inverse Ising problem it was first used by Ravikumar *et al* [27], albeit in the context of graph sign-sparsity reconstruction; two of us showed recently that it outperforms many other approximate inverse Ising schemes on the Sherrington-Kirkpatrick model, and in several other examples [23]. Although this paper is the first use of the pseudo-likelihood maximization method in DCA, the idea to use pseudo-likelihoods for PSP is not completely novel. Balakrishnan *et al.* [8] devised a version of this idea, but using a set up rather different from that of [2], regarding *e.g.* what portions of the data bases and which measures of prediction accuracy were used, and not also couched in the language of inverse statistical mechanics. While a competitive evaluation between [2] and [8] is still open, we have not attempted to do so in this work.

Other ways of deducing direct interactions in PSP, not motivated from the Potts model but in somewhat similar probabilistic settings have been suggested in the last few years. A fast method utilizing Bayesian networks was provided by Burger and van Nimwegen [7]. More recently Jones *et al.* [9] introduced a procedure called *PSICOV* (Protein Sparse Inverse COVariance). While DCA and PSICOV both appear capable of outperforming the Bayesian network approach [2, 9], their relative efficiency is currently open to investigation, and has not been assessed in this work.

Finally, predicting amino acid contacts is not only a goal in itself, it is also a means to assemble protein complexes [40, 41] and to predict full 3D protein structures [3, 4, 42]. Such tasks require additional work, using the DCA results as input, and are outside the scope of the present paper.

III. METHOD DEVELOPMENT

A. The Potts model

Let $\boldsymbol{\sigma} = (\sigma_1, \sigma_2, \cdots, \sigma_N)$ represent the amino acid sequence of a domain with length N. Each σ_i takes on values in $\{1, 2, ..., q\}$, with $q = 21$: one state for each of the 20 naturally occurring amino acids and one additional state to represent gaps. Thus, an MSA with B aligned sequences from a domain family can be written as an integer array $\{\boldsymbol{\sigma}^{(b)}\}_{b=1}^{B}$, with one row per sequence and one column per chain position. Given an MSA, the empirical individual and pairwise *frequencies* can be calculated as

$$\begin{aligned} f_i(k) &= \frac{1}{B}\sum_{b=1}^{B}\delta(\sigma_i^{(b)}, k), \\ f_{ij}(k,l) &= \frac{1}{B}\sum_{b=1}^{B}\delta(\sigma_i^{(b)}, k)\ \delta(\sigma_j^{(b)}, l). \end{aligned} \tag{4}$$

where $\delta(a, b)$ is the Kronecker symbol taking value one if both arguments are equal, and zero else. $f_i(k)$ is hence the fraction of sequences for which the entry on position i is amino acid k, gaps counted as a 21st amino acid. Similarly, $f_{ij}(k, l)$ is the fraction of sequences in which the position pair (i, j) holds the amino acid combination (k, l). Connected correlations are given as

$$c_{ij}(k,l) = f_{ij}(k,l) - f_i(k)\, f_j(l). \tag{5}$$

A (generalized) Potts model is the simplest probabilistic model $P(\boldsymbol{\sigma})$ which can reproduce the empirically observed $f_i(k)$ and $f_{ij}(k, l)$. In analogy to (1) it is defined as

$$P(\boldsymbol{\sigma}) = \frac{1}{Z}\exp\left(\sum_{i=1}^{N} h_i(\sigma_i) + \sum_{1\le i<j\le N} J_{ij}(\sigma_i,\sigma_j)\right), \tag{6}$$

in which $h_i(\sigma_i)$ and $J_{ij}(\sigma_i, \sigma_j)$ are parameters to be determined through the constraints

$$\begin{aligned} P(\sigma_i = k) &= \sum_{\substack{\boldsymbol{\sigma} \\ \sigma_i=k}} P(\boldsymbol{\sigma}) = f_i(k), \\ P(\sigma_i = k, \sigma_j = l) &= \sum_{\substack{\boldsymbol{\sigma} \\ \sigma_j=l \\ \sigma_i=k}} P(\boldsymbol{\sigma}) = f_{ij}(k,l), \end{aligned} \tag{7}$$

It is immediate that the probabilistic model, which maximizes entropy while satisfying Eq. (7), must take the Potts model form. Finding a Potts model which matches empirical frequencies and correlations is therefore referred to as a *maximum entropy inference*. On the other hand, Eq. (6) – with parameters to be adjusted to data – is in itself a valid inference problem on a well-defined model class, and this is the perspective which will be used here. We note that the Ising and the Potts models (and most models which would normally be considered in statistical mechanics) are examples of *exponential families*, which have the property that means and correlations are *sufficient statistics* [43–45]. Given unlimited computing power to determine Z, reconstruction can not be done better using all the data compared to using only (empirical) means and (empirical) correlations. It is only when one cannot compute Z exactly and have to resort to approximate methods, that using directly all the data can bring any advantage.

B. Model parameters and gauge invariance

The total number of parameters of Eq, (6) is $Nq + \frac{N(N-1)}{2}q^2$, but, in fact, the model as it stands is overparameterized in the sense that distinct parameter sets can describe the same probability distribution. It is easy to see that the number of non-redundant parameters is $N(q-1) + \frac{N(N-1)}{2}(q-1)^2$, cf. an Ising model($q = 2$), which has $\frac{N(N+1)}{2}$ parameters if written as in Eq. (1) but would have $2N^2$ parameters if written in the form of Eq. (6).

A gauge choice for the Potts model, which eliminates the overparametrization in a similar manner as in the Ising model (and reduces to that case for $q = 2$), is

$$\sum_{s=1}^{q} J_{ij}(k,s) = \sum_{s=1}^{q} J_{ij}(s,l) = \sum_{s=1}^{q} h_i(s) = 0, \tag{8}$$

for all i, j, k, and l. In the PSP context the last index ($i = q$, which stands for the gap in an alignment) is special, we can therefore chose a gauge where all interaction energies are measured with respect to this value, i.e.

$$J_{ij}(q,l) = J_{ij}(k,q) = h_i(q) = 0, \tag{9}$$

for all i, j, k, and l, cf. [2]. This gauge choice corresponds to a lattice gas model with $q - 1$ different particle types, and a maximum occupation number one.

Using either (8) or (9) reconstruction is well-defined, and it is straight-forward to translate results obtained in one gauge to the other.

C. The inverse Potts problem

Given a set of independent equilibrium generations $\{\boldsymbol{\sigma}^{(b)}\}_{b=1}^{B}$ of the model Eq. (6), the ordinary statistical approach to inferring parameters $\{\mathbf{h}, \mathbf{J}\}$ would be to let those parameters maximize the likelihood (i.e. the probability of generating the data set for a given set of parameters). This is equivalent to minimizing the (rescaled) negative log-likelihood function

$$nll = -\frac{1}{B}\sum_{b=1}^{B}\log P(\boldsymbol{\sigma}^{(b)}). \tag{10}$$

For the Potts model (6), this becomes

$$\begin{aligned} nll(\mathbf{h},\mathbf{J}) \;&=\; \log Z - \sum_{i=1}^{N}\sum_{k=1}^{q} f_i(k)h_i(k) \\ &- \sum_{1\le i<j\le N}\;\sum_{k,l=1}^{q} f_{ij}(k,l)J_{ij}(k,l). \end{aligned} \tag{11}$$

nll is differentiable, so minimizing it means looking for a point at which $\partial_{h_i(k)} nll = 0$ and $\partial_{J_{ij}(k,l)} nll = 0$. Hence, ML estimates will satisfy

$$\begin{aligned} \partial_{h_i(k)} \log Z - f_i(k) \;&=\; 0, \\ \partial_{J_{ij}(k,l)} \log Z - f_{ij}(k,l) \;&=\; 0. \end{aligned} \tag{12}$$

To achieve this maximization computationally, we need to be able to calculate the partition function Z of Eq. (6) for any realization of the parameters $\{\mathbf{h},\mathbf{J}\}$. This problem is computationally intractable for any reasonable systems size. Approximate maximization is essential, and we will show that even relatively simple approximation schemes lead to accurate PSP results.

D. Naive mean-field inversion

The mfDCA algorithm in [2] is based on the simplest, and computationally most efficient approximation, i.e. *naive mean-field inversion.* It starts from the proper generalization of (2), *cf.* [46], and then uses linear response: The J's in the lattice-gas gauge Eq. (9) become:

$$J^{NMFI}_{ij,kl} = -(\mathbf{C}^{-1})_{ab}, \tag{13}$$

where $a = (q-1)(i-1)+k$ and $b = (q-1)(j-1)+l$. The matrix $\mathbf{C}$ is the $N(q-1)\times N(q-1)$ covariance matrix assembled by joining the $N(q-1)\times N(q-1)$ values $c_{ij}(k,l)$ as defined in Eq. (5), but leaving out the last state q. In Eq. (13), $i,j\in\{1,...,N\}$ are site indices, and k,l run from 1 to $q-1$. Under gauge Eq. (9), all the other coupling parameters are zero. The term "naive" has become customary in the inverse statistical mechanics literature, often used to highlight the difference to a Thouless-Anderson-Palmer level inversion or one based on the Bethe approximation. The original meaning of this term lies, as far as we are aware, in Information Geometry [47, 48].

E. Pseudo-likelihood maximization

Pseudo-likelihood substitutes the likelihood (11) by the conditional probability of observing one variable σ_r given observations of all the other variables $\boldsymbol{\sigma}_{\setminus r}$. That is, the starting point is

$$\begin{aligned} P(\sigma_r \;&=\; \sigma_r^{(b)}|\boldsymbol{\sigma}_{\setminus r} = \boldsymbol{\sigma}_{\setminus r}^{(b)}) \\ &=\; \frac{\exp\left(h_r(\sigma_r^{(b)}) + \sum\limits_{\substack{i=1\\ i\neq r}}^{N} J_{ri}(\sigma_r^{(b)},\sigma_i^{(b)})\right)}{\sum_{l=1}^{q}\exp\left(h_r(l) + \sum\limits_{\substack{i=1\\ i\neq r}}^{N} J_{ri}(l,\sigma_i^{(b)})\right)}, \end{aligned} \tag{14}$$

where, for notational convenience, we take $J_{ri}(l,k)$ to mean $J_{ir}(k,l)$ when $i<r$. Given an MSA, we can maximize the conditional likelihood by minimizing

$$g_r(\mathbf{h}_r,\mathbf{J}_r) = -\frac{1}{B}\sum_{b=1}^{B}\log\left[P_{\{\mathbf{h}_r,\mathbf{J}_r\}}(\sigma_r = \sigma_r^{(b)}|\boldsymbol{\sigma}_{\setminus r} = \boldsymbol{\sigma}_{\setminus r}^{(b)})\right]. \tag{15}$$

Note that this only depends on $\mathbf{h}_r$ and $\mathbf{J}_r = \{\mathbf{J}_{ir}\}_{i\neq r}$, that is, on the parameters featuring node r. If (15) is used for all r this leads to unique values for the parameters $\boldsymbol{h_r}$ but typically different predictions for $\boldsymbol{J_{rq}}$ and $\boldsymbol{J_{qr}}$ (which should be the same). Maximizing (15) must therefore be supplemented by some procedure on how to reconcile different values of $\boldsymbol{J_{rq}}$ and $\boldsymbol{J_{qr}}$; one way would be to simply use their average $\frac{\boldsymbol{J_{rq}}+\boldsymbol{J_{qr}}}{2}$ [27].

We here reconcile different $\boldsymbol{J_{rq}}$ and $\boldsymbol{J_{qr}}$ by maximizing an objective function by adding f_r for all nodes:

$$\begin{aligned} npll(\mathbf{h},\mathbf{J}) &= \sum_{r=1}^{N} g_r(\mathbf{h}_r,\mathbf{J}_r) \\ &= -\frac{1}{B}\sum_{r=1}^{N}\sum_{b=1}^{B}\log\left[P_{\{\mathbf{h}_r,\mathbf{J}_r\}}(\sigma_r = \sigma_r^{(b)}|\boldsymbol{\sigma}_{\setminus r} = \boldsymbol{\sigma}_{\setminus r}^{(b)})\right]. \end{aligned} \tag{16}$$

where the abbreviation *npll* stands for *negative pseudo-log-likelihood.* Minimizers of *npll* generally do not minimize *nll*; the replacement of likelihood with pseudo-likelihood alters the outcome. Note however, that replacing *nll* by *npll* resolves the computational intractability of the parameter optimization problem: instead of depending on the full partition function, the normalization of the conditional probability (14) contains only a single summation over the $q = 21$ Potts states. The intractable average over the $N-1$ conditioning spin variables is replaced by an empirical average over the data set in Eq. (16).

F. Regularization

A Potts model describing a protein family with sequences of 50-300 amino acids requires ca. $5\cdot 10^5 - 2\cdot 10^7$ parameters. At present, few protein families are in this range in size, and *regularization* is therefore needed to avoid over-fitting. In naive mean-field inversion the problem results in an empirical covariance matrix, which typically not of full rank, and Eq. (13) is not well-defined.

In [2], one of the authors therefore used the pseudocount method where frequencies and empirical correlations are adjusted using a regularization variable λ:

$$\begin{aligned} f_i(k) &= \frac{1}{\lambda+B}\left[\frac{\lambda}{q}+\sum_{b=1}^{B}\delta(\sigma_i^{(b)},k)\right], \qquad (17)\\ f_{ij}(k,l) &= \frac{1}{\lambda+B}\left[\frac{\lambda}{q^2}+\sum_{b=1}^{B}\delta(\sigma_i^{(b)},k)\;\delta(\sigma_j^{(b)},l)\right]. \end{aligned}$$

The pseudocount is a proxy for many observations, which – if they would exist – would increase the rank of the correlation matrix; the pseudo-count method hence promotes invertibility of the matrix in Eq. (13). It was observed in [2] that for good performance in DCA, the pseudocount parameter λ has to be taken fairly large, on the order of B.

In the pseudo-likelihood maximization method, we take the standard route of adding a penalty term to the objective function:

$$\{\mathbf{h}^{PLM},\mathbf{J}^{PLM}\} = \underset{\{\mathbf{h},\mathbf{J}\}}{\operatorname{argmin}}\{npll(\mathbf{h},\mathbf{J})+R(\mathbf{h},\mathbf{J})\}. \qquad (18)$$

The turnout is then a trade-off between likelihood maximization and whatever qualities R is pushing for. Ravikumar *et al.* [27] pioneered the use of l_1 regularizers for the inverse Ising problem, which forces a finite fraction of parameters to assume value zero, thus effectively reducing the number of parameters. This approach is not appropriate here since we are concerned with the accuracy of the strongest predicted couplings which would typically be distorted by an l_1 penalty; for our purposes it makes no substantial difference if weak couplings are inferred to be small or set precisely to zero. Our choice for R is therefore the simpler l_2 norm

$$R_{l_2}(\mathbf{h},\mathbf{J}) = \lambda_h\sum_{r=1}^{N}||\mathbf{h}_r||_2^2+\lambda_J\sum_{i=1}^{N-1}\sum_{j=i+1}^{N}||\mathbf{J}_{ij}||_2^2. \qquad (19)$$

using two regularization parameters λ_h and λ_J for field and coupling parameters. An advantage of a regularizer is that it eliminates the need to fix a gauge, since among all parameter sets related by a gauge transformation, i.e. all parameter sets resulting in the same Potts model, there will be exactly one set which minimizes the strictly convex regularizer. Actually, for the case of the l_2 norm, it can be shown easily that this leads to the Ising-type gauge condition, Eq. (8).

To summarize this discussion: For NMFI, we regularize with pseudocounts under the gauge constraints Eq. (9). For PLM, we regularize with R_{l_2} under the full parametrization.

G. Sequence re-weighting

Maximum-likelihood inference of Potts models relies – as discussed above – on the assumption that the B sample configurations in our data set are independently generated from Eq. (6). This assumption is not exactly true for biological sequence data which have a *phylogenetic* bias. In particular, in the data bases there are many protein sequences from related species which did not have enough time of independent evolution to reach statistical independence. Furthermore, the selection of sequenced species in the genomic databases is dictated by human interest, and not by the aim to have an as independent as possible sampling in the space of all functional amino-acid sequences. A way to mitigate effects of uneven sampling, employed in [2], is to equip each sequence $\boldsymbol{\sigma}^{(b)}$ with a *weight* w_b which regulates its impact on the parameter estimates. Sequences deemed unworthy of independent-sample status (too similar to other sequences) can then have their weight lowered, whereas sequences, which are quite different from all other sequences, will contribute with a higher weight to the amino-acid statistics.

A simple but efficient way (cf. [2]) is to measure the similarity $\mathrm{sim}(\boldsymbol{\sigma}^{(a)},\boldsymbol{\sigma}^{(b)})$ of two sequences $\boldsymbol{\sigma}^{(a)}$ and $\boldsymbol{\sigma}^{(b)}$ as the fraction of conserved positions (*i.e.* identical amino acids), and compare this fraction to a pre-selected threshold x , $0<x<1$. The weight given to a sequence $\boldsymbol{\sigma}^{(b)}$ can then be set to $w_b=\frac{1}{m_b}$, where m_b was the number of sequences in the MSA similar to $\boldsymbol{\sigma}^{(b)}$:

$$m_b = |\{a\in\{1,...,B\}:\mathrm{sim}(\boldsymbol{\sigma}^{(a)},\boldsymbol{\sigma}^{(b)})\geq x\}|. \qquad (20)$$

In [2], a suitable threshold x was found to be 0.8, results only weakly dependent on this choice throughout $0.7<x<0.9$. We have here followed the same procedure using threshold $x=0.9$. The corresponding re-weighted frequency counts then become

$$\begin{aligned} f_i(k) &= \frac{1}{\lambda+B_{eff}}\left[\frac{\lambda}{q}+\sum_{b=1}^{B}w_b\,\delta(\sigma_i^{(b)},k)\right], \qquad (21)\\ f_{ij}(k,l) &= \frac{1}{\lambda+B_{eff}}\left[\frac{\lambda}{q^2}+\sum_{b=1}^{B}w_b\,\delta(\sigma_i^{(b)},k)\,\delta(\sigma_j^{(b)},l)\right], \end{aligned}$$

where $B_{eff}=\sum_{b=1}^{B}w_b$ becomes a measure of the number of effectively non-redundant sequences.

In the pseudo-likelihood we use the direct analogue of Eq. (21), *i.e.*

$$\begin{aligned} &npll(\mathbf{h},\mathbf{J}) \qquad (22)\\ &= -\frac{1}{B_{eff}}\sum_{b=1}^{B}w_b\sum_{r=1}^{N}\log\left[P_{\{\mathbf{h}_r,\mathbf{J}_r\}}(\sigma_r=\sigma_r^{(b)}|\boldsymbol{\sigma}_{\backslash r}=\boldsymbol{\sigma}_{\backslash r}^{(b)})\right]. \end{aligned}$$

As in the frequency counts, each sequence is considered to contribute a weight w_b, instead of the standard weight one used in i.i.d. samples.

H. Interaction scores

In the inverse Ising problem each interaction is scored by one scalar coupling strength J_{ij}. These can easily be ordered, *e.g.* by absolute size. In the inverse Potts problem, each each position pair (i, j) is characterized by a whole $(q-1)\times(q-1)$ matrix $\mathbf{J}_{ij}$, and some scalar score is needed in order to evaluate the 'coupling strength' of two sites.

In [1] and [2] the score used is the *direct information* (DI), *i.e.* the mutual information of a restricted probability model not including any indirect coupling effects between the two positions to be scored. The construction of DI goes as follows: For each position pair (i, j), (the estimate of) $\mathbf{J}_{ij}$ is used to set up a 'direct distribution' involving only nodes i and j,

$$P^{(dir)}_{ij}(k,l) \sim \exp\left(J_{ij}(k,l) + h'_{i,k} + h'_{j,l}\right). \tag{23}$$

$h'_{i,k}$ and $h'_{j,l}$ are new fields, computed as to ensure agreement of the marginal single-site distributions with the empirical individual frequency counts $f_i(k)$ and $f_j(l)$. The DI score is now calculated as the mutual information of $P^{(dir)}$:

$$DI_{ij} = \sum_{k,l=1}^{q} P^{(dir)}_{ij}(k,l)\log\left(\frac{P^{(dir)}_{ij}(k,l)}{f_i(k)\,f_j(l)}\right). \tag{24}$$

A nice characteristics of DI is its invariance with respect to the gauge freedom of the Potts model, *i.e.* both choices Eqs. (8) and (9) (or any other valid choice) generate identical DI.

In the pseudo-likelihood approach, we prefer not to use DI, as this would require a pseudocount λ to regularize the frequencies in the DI computation, introducing a third regularization variable in addition to λ_h and λ_J. Another possible scoring function, already mentioned but not used in [1], is the *Frobenius norm*

$$\|\mathbf{J}_{ij}\|_2 = \sqrt{\sum_{k,l=1}^{q} J_{ij}(k,l)^2}. \tag{25}$$

Unlike DI, (25) is *not* independent of gauge choice, so one must be a bit careful. As was noted in [1], the zero sum gauge (8) minimizes the Frobenius norm, in a sense making (8) the most appropriate one for the score (25). Recall from above that our pseudo-likelihood uses the full representation and fixes the gauge by the regularization terms R_{l_2}. Our procedure is therefore first to infer the interaction parameters using using the pseudo-likelihood and the regularization, and then change to the zero-sum gauge:

$$J'_{ij}(k,l) = J_{ij}(k,l) - J_{ij}(\cdot,l) - J_{ij}(k,\cdot) + J_{ij}(\cdot,\cdot), \tag{26}$$

where '·' denotes average over the concerned position. One can show that (26) preserves the probabilities of (6) (after altering the fields appropriately) and that $J'_{ij}(k,l)$ satisfy (8). A possible Frobenius norm score is hence

$$FN_{ij} = \|\mathbf{J}'_{ij}\|_2 = \sqrt{\sum_{k,l=1}^{q} J'_{ij}(k,l)^2}. \tag{27}$$

Lastly we borrow an idea from Jones *et al.* [9], whose PSICOV method also used a norm rank (l_1-norm instead of Frobenius norm), but scores are adjusted by an *average product correction* (APC) term. APC was introduced in [49] to suppress effects from phylogenetic bias and insufficient sampling. Incorporating also this correction, we have our scoring function

$$CN_{ij} = FN_{ij} - \frac{FN_{\cdot j}FN_{i\cdot}}{FN_{\cdot\cdot}}, \tag{28}$$

where CN stands for 'corrected norm'.

IV. EVALUATING THE PERFORMANCE OF MFDCA AND PLMDCA ACROSS PROTEIN FAMILIES

We have performed numerical experiments using mfDCA and plmDCA on a number of domain families from the Pfam database; here we report and discuss these results.

A. Domain families, native structures, and true-positive rates

The speed of mfDCA enabled Morcos et al. [2] to conduct a large-scale analysis using 131 families. PLM is computationally more demanding than NMFI, so we chose to start with a smaller collection of 17 families, listed in Table I. To ease the numerical effort, we chose families with relatively small N.

To reliably assess how good a contact prediction is, something to regard as a "gold standard" is helpful. For each of the 17 families we have therefore selected one representative high-resolution X-ray crystal structure (resolution below 3Å), see the last column of Table I for the corresponding PDB identification.

From these native protein structures, we have extracted position-position distances $d(i, j)$ for each pair of sequence positions, by measuring the minimal distance between any two heavy atoms belonging to the amino acids present in these positions. Fig. 2.a shows the distribution of these distances in all considered families. Three peaks protrude from the background distribution: One at small distances below 1.5Å, a second at about 3-5Å and the third at about 7-8Å. The first peak corresponds to the peptide bonds between sequence neighbors, whereas the other two peaks correspond to non-trivial contacts between amino acids, which may be distant along the protein backbone, as can be seen from panels b and c of Fig. 2, which collect only distances between positions i and j with minimal separation $|i-j| \geq 5$ resp. $|i-j| \geq 15$. Following [2], we take the peak at 3-5Å to presumably correspond to short-range interactions like hydrogen bonds or secondary-structure contacts, whereas the last peak likely corresponds to long-range, possibly water-mediated interactions. These peaks contain the

Family ID	N	B	B_{eff} (90%)	PDB ID
PF00011	102	7151	3481	2bol
PF00013	58	11484	3785	1wvn
PF00014	53	3090	1812	5pti
PF00017	77	4403	1741	1o47
PF00018	48	8993	3354	2hda
PF00027	91	17830	9036	3fhi
PF00028	93	18808	8317	2o72
PF00035	67	5584	2254	1o0w
PF00041	85	26172	10631	1bqu
PF00043	95	9619	5141	6gsu
PF00046	57	15445	3314	2vi6
PF00076	70	31837	14125	1g2e
PF00081	82	5867	1510	3bfr
PF00084	56	9816	4345	1elv
PF00105	70	4842	1277	1gdc
PF00107	130	28022	12114	1a71
PF00111	78	11941	5805	1a70

TABLE I: Domain families included in our study, listed with Pfam ID, length N, number of sequences B, the number of effective sequences B_{eff} (under 90% re-weighting), and the PDB structure used to access the DCA prediction quality.

non-trivial information we would like to extract from sequence data using DCA. In order to accept the full second peak, we have chosen a distance cutoff of 8.5Å for true contacts, slightly larger than the value of 8Å used in [2].

Accuracy results are here reported primarily using *true-positive* (TP) rates, also the principal measurement of in [2] and [9]. The TP rate for p is the fraction of the p strongest-scored pairs which are actually contacts in the crystal structure, defined as described above. To exemplify TP rates, let us jump ahead and look at Fig. 3. For PLM and protein family PF00076, the TP rate is one up to $p = 80$, which means that all 80 highest-CN pairs are genuine contacts in the crystal structure. At $p = 200$, the TP rate has dropped to 0.78, so $0.78 \cdot 200 = 156$ of the top 200 highest-CN pairs are contacts, while 44 are not.

B. Parameter settings

To set the stage for comparison, we started by running initial trials on 17 families using both NMFI and PLM with many different regularization and re-weighting strengths. Re-weighting indeed raised the TP rates, and, as was reported in [2] for the 131 families, results seemed robust toward the exact choice of the limit x around $0.7 \leq x \leq 0.9$. We chose $x = 0.9$ to use throughout the study.

In what follows, NMFI results are reported using the same list of pseudocounts as in Fig. S11 in [2]: $\lambda = w \cdot B_{eff}$ with $w = \{0.11, 0.25, 0.43, 0.67, 1.0, 1.5, 2.3, 4.0, 9.0\}$. During our analysis we also ran intermediate values, and we found this covering to be sufficiently dense. We give outputs from two versions of NMFI: NMFI-DI and NMFI-DI(true). The former uses pseudocounts for all calculations, whereas the latter switches to true frequencies when it gets to the evaluations of the DI scores.

With l_2-regularization in the PLM algorithm, outcomes were robust against the precise choice of λ_h; TP rates were almost identical when λ_h was changed between 0.001 and 0.1. We therefore chose $\lambda_h = 0.01$ for all experiments. What mattered, rather, was the coupling regularization parameters λ_J, for which we did a systematic scan from $\lambda_J = 0$ and up using step-size 0.005.

So, to summarize, the results reported here are based on $x = 0.9$, cutoff 8.5Å, and $\lambda_h = 0.01$, and λ and λ_J drawn from collections of values as described above.

C. Main comparison of mfDCA and plmDCA

Fig. 3 shows TP rates for the different families and methods. We see that TP of plmDCA (PLM) rates are consistently greater than the ones of mfDCA (NMFI), especially for families with large B_{eff}. For what concerns the two NMFI versions: NMFI-DI(true) avoids the strong failure seen in NMFI-DI for PF00084, but for most other families, see in particular PF00014 and PF00081, the performance instead drops using marginals without pseudo-counts in the DI calculation. For both NMFI-DI and NMFI-DI(true), the best regularization was found to be $\lambda = 1 \cdot B_{eff}$, the same value as used in [2]). For PLM, the best parameter choice was $\lambda_J = 0.01$. Interestingly, this same regularization parameter were optimal for basically all families. This is somewhat surprising, since both N and B_{eff} span quite wide ranges (48-130 and 1277-14225 respectively).

In the following discussion, we leave out all results for NMFI-DI(true) and focus on PLM vs. NMFI-DI, i.e. the version used in [2]). All plots remaining in this section use the optimal regularization values: $\lambda = B_{eff}$ for NMFI and $\lambda_J = 0.01$ for PLM.

TP rates only classify pairs as contacts ($d(i,j) < 8.5$Å) or non-contacts ($d(i,j) \geq 8.5$Å). To give a more detailed view of how scores correlate with spatial separation, we show in Fig. 4 a scatter plot of the score vs. distance for all pairs in all families. PLM and NMFI-DI both manage to detect the peaks seen in the true distance distribution of Fig. 2.a, in the sense that high scores are observed almost exclusively at distances below 8.5Å. Both methods agree that interactions get, on average, progressively weaker going from peak one, to two, to three, and finally to the bulk. We note that the dots scatter differently across the PLM and NMFI-DI figures, reflecting the two separate scoring techniques: DI are strictly non-negative, whereas APC corrected norms can assume negative values. We also observe how sparse the extracted signal is: most spatially close pairs do not show elevated scores.

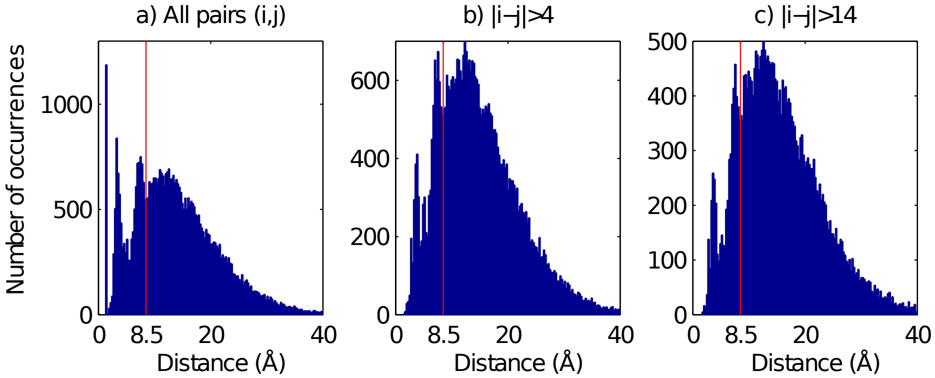


FIG. 2: Histograms of crystal-structure distances pooled from all 17 families. The headers state the types of pairs included. The red line is our contact cutoff 8.5Å.

However, from the other side almost all strongly coupled pairs are close, so the biological hypothesis of Sec. II is well supported here.

Fig. 5 shows scatter plots of scores for PLM and NMFI-DI for some selected families. Qualitatively the same patterns were observed for all families. The points are clearly correlated, so, to some extent, PLM and NMFI-DI agree on the interaction strengths. Due to the different scoring schemes, we would not expect numerical coincidence of scores. Many of PLM's top-scoring position pairs have also top scores for NMFI-DI and vice versa. The largest discrepancy is in how much stronger NMFI-DI responds to pairs with small $|i-j|$; the blue crosses tend to shoot out to the right. PLM agrees that many of these neighbor pairs interact strongly, but, unlike NMFI-DI, it also shows rivaling strengths for many $|i-j|>4$-pairs.

An even more detailed picture is given by considering *contact maps*, see Fig. 7. The tendency observed in the last scatter plots remains: NMFI-DI has a larger portion of highly scored pairs in the neighbor zone, which are the middle stretches in these figures. An important observation is, however, that clusters of contacting pairs with long 1D sequence separation are captured by both algorithms.

In summary, the results suggest that the PLM method offers some interesting progress compared to NMFI. However, let us also note that in the comparison we had also to change both scoring and regularization styles. It is thus conceivable that a naive mean-field inversion with the new scoring and regularization could be more competitive with PLM. Indeed, upon further investigation, detailed in Appendix B, we found that part of the improvement in fact does stem from the new score. In the comparison to follow, we therefore add results from NMFI-CN, an updated version of the code used in [2] which scores by CN instead of DI.

D. Run times

In general, NMFI, which is merely a matrix inversion, is very quick compared with PLM; most families in this study took only seconds to run through the NMFI code.

In contrast to message-passing based method used in [1], a DCA using PLM is nevertheless feasible for all protein families in PFAM. The objective function in PLM is a sum over nodes and samples and its execution time is therefore expected to depend both on B (number of members of a protein family in PFAM) and N (length of the aligned sequences in a protein family). Since B varies over a larger range than N dependance is further expected to be mainly on B.

On one core of a standard desktop computer, using a MATLAB-interfaced C-implementation of conjugate gradient (CG) descent, run times for PF00014, PF00017, and PF00018 (small N and B) were 9, 22, and 12 minutes respectively. For PF00041 (small N but larger B) one run took 2.5 hours. For larger values of N in the set explored run times grow approximately apace. For example, running times for PF00026 ($N=314$) and PF00006 ($N=215$) were 15 and 12 hours respectively.

All of these times were obtained cold-starting with all fields and couplings at zero. Presumably one can improve by using an appropriate initial guess obtained, say, from NMFI. This has however not been implemented here. Also, by minimizing each f_r separately, PLM would have amounted to N separate (multi-class) logistic regression problems, completely open to parallel solving. This variant of PLM has however not been used in this work. Finally, we note that if computational speed is an issue then for large families one could also have chosen to use only a subset of the sequences, effectively lowering B.

V. DISCUSSION

In this work, we have shown that a direct-couping analysis built on pseudo-likelihood maximization (plmDCA)

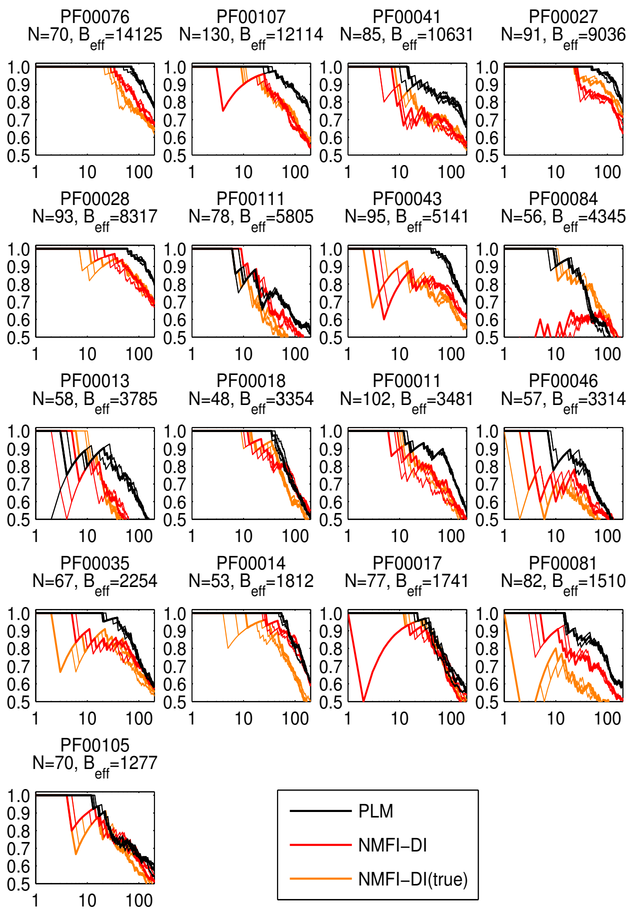


FIG. 3: Contact-detection results for the 17 families, sorted by B_{eff}. Y-axes are TP rates and x-axes are the number of predicted contacts p, based on pairs with $|i-j|>4$. The three curves for each method are the three regularization levels yielding highest TP rates across all families. The thickened curve highlights the best one out of these three ($\lambda = B_{eff}$ for NMFI and $\lambda_J = 0.01$ for PLM).

consistently outperforms the previously described mean-field based analysis (mfDCA), as assessed across a number of large protein-domain families. The advantage of the pseudo-likelihood approach was found to be partially intrinsic, and partly contingent on using a sampling-corrected Frobenius norm to score inferred direct statistical coupling matrices.

On one hand, this improvement might not be surprising: it is known that, for very large data sets, pseudo-likelihood maximization becomes asymptotically equivalent to full maximum-likelihood inference, whereas mean-field inference remains intrinsically approximate, and this may result in an improved PLM performance also for finite data sets [23].

On the other hand, the above advantage holds if and only if the following two conditions are fulfilled: data a drawn independently from a probability distribution, and this probability distribution is the the Boltzmann distribution of a Potts model. None of these two conditions actually hold for real protein sequences. On artificial data, also refined mean-field methods (Thouless-Anderson-Palmer equations, Bethe approximation) lead to improved model inference as compared to naive mean-field inversion, cf. *e.g.* [14, 16, 17, 21], but no such improvement has been observed in real protein data [2]. The results of the paper are therefore interesting and highly non-trivial. They also suggest that other model-learning methods from statistics such as “Contrastive Divergence” [50] or the more recent “Noise-Contrastive Estimation” [51], could be explored to further increase our capacity to extract structural information from protein sequence data.

Disregarding the improvements, we find that overall the predicted contact pairs for plmDCA and mfDCA are highly overlapping, illustrating the robustness of DCA results with respect to the algorithmic implementation. This observations suggests that, in the context of modeling the sequence statistics by pairwise Potts models, most extractable information might already already be extracted from the MSA. However, it may well also be that there is alternative information hidden in the sequences, for which we would need to go beyond pair-wise models, or integrate the physico-chemical properties of different amino acids into the procedure, or extract even more information from large sets of evolutionarily related amino-acid sequences. DCA is only a step into this direction.

In our work we have seen, that simple sampling corrections, more precisely sequence-re-weighting and the average-product correction of interaction scores, lead to an increased accuracy in predicting 3D contacts of amino-acids, which are distant on the protein’s backbone. It is, however, clear that these somewhat heuristic statistical fixes cannot correct for the complicated hierarchical phylogenetic relationships between proteins, and that more sophisticated methods would be needed to disentangle phylogenetic from functional correlations in massive sequence data. To do so is an open challenge, which would leave the field of *equilibrium* inverse statistical mechanics, but where methods on inverse statistical mecahnics may still play a useful role.

Acknowledgments

This work has been supported by the Academy of Finland as part of its Finland Distinguished Professor program, project 129024/Aurell, and through the Center of Excellence COIN. Discussions with S. Cocco and R. Monasson are gratefully acknowledged.

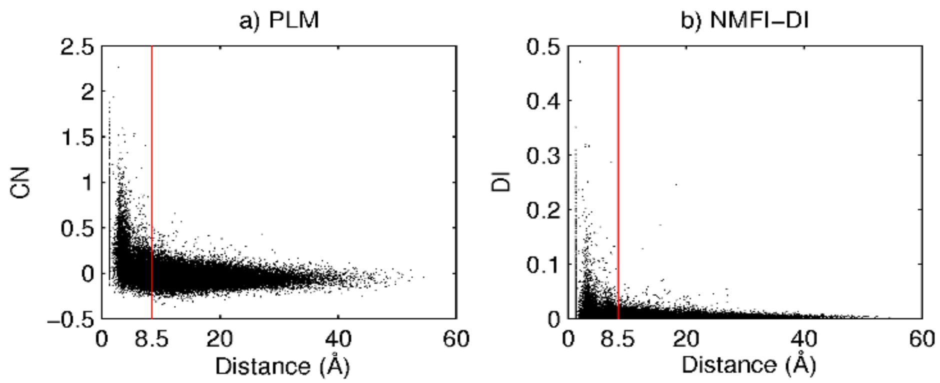


FIG. 4: Score plotted against distance for all position pairs in all 17 families. The red line is our contact cutoff at 8.5Å.

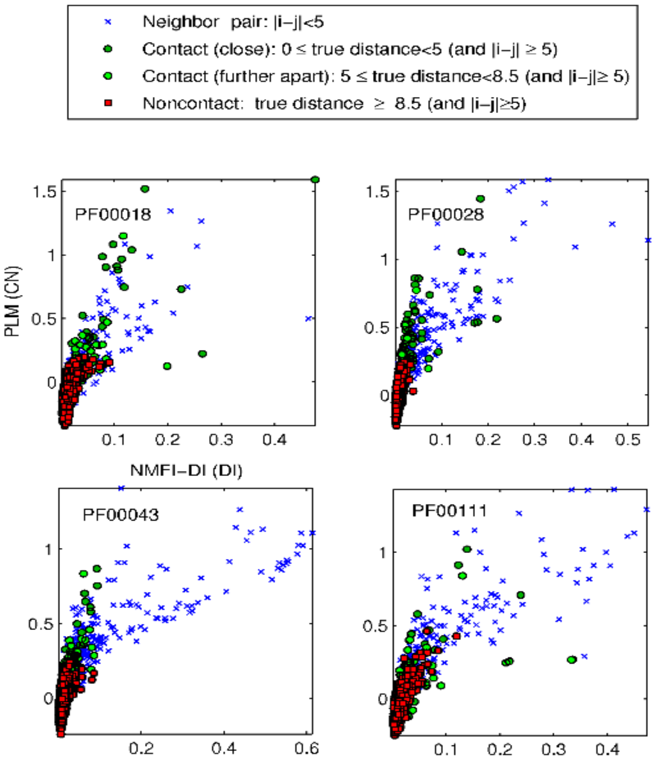


FIG. 5: Scatter plots of interaction scores for PLM and NMFI-DI from four families. For all plots, the axes are as indicated by the top left one. The distance unit in the top box is Å.

[1] M. Weigt, R. White, H. Szurmant, J. Hoch, and T. Hwa, Proc. Natl. Acad. Sci. U. S. A. **106**, 67 (2009).

[2] F. Morcos, A. Pagnani, B. Lunt, A. Bertolino, D. S. Marks, C. Sander, R. Zecchina, J. N. Onuchic, T. Hwa, and M. Weigt, Proc. Natl. Acad. Sci. U. S. A. **108**, E1293 (2011).

[3] D. S. Marks, L. J. Colwell, R. P. Sheridan, T. A. Hopf, A. Pagnani, R. Zecchina, and C. Sander, arXiv:1110.5091 (2011).

[4] J. Sulkowska, F. Morcos, M. Weigt, T. Hwa, , and J. Onuchic, Proc. Natl. Acad. Sci. **109**, 10340 (2012).

[5] T. Nugent and D. T. Jones, Proceedings of the National

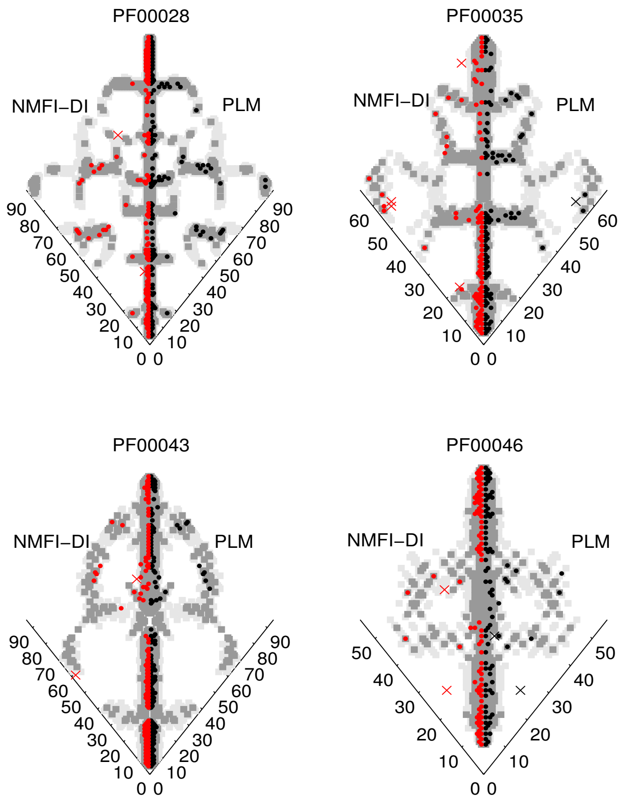


FIG. 6: Predicted contact maps for PLM and NMFI-DI for four families. A pair (i,j)'s placement in the plots is found by matching positions i and j on the axes. Contacts are indicated in gray. True and false positives are represented by circles and crosses, respectively. Each figure shows the $1.5N$ strongest ranked pairs (including neighbors) for that family.

Academy of Sciences **109**, E1540 (2012).

[6] M. Punta, P. C. Coggill, R. Y. Eberhardt, J. Mistry, J. G. Tate, C. Boursnell, N. Pang, K. Forslund, G. Ceric, J. Clements, et al., Nucleic Acids Res. **40**, D290 (2012).

[7] L. Burger and E. van Nimwegen, PLoS Comput. Biol. **6**, E1000633 (2010).

[8] S. Balakrishnan, H. Kamisetty, J. Carbonell, S. Lee, and C. Langmead, Proteins: Struct., Funct., Bioinf. **79**, 1061 (2011).

[9] D. T. Jones, D. W. A. Buchan, D. Cozzetto, and M. Pontil, Bioinformatics **28**, 184 (2012).

[10] H. J. Kappen and F. B. Rodriguez, in *Advances in Neural Information Processing Systems* (The MIT Press, 1998), pp. 280–286.

[11] E. Schneidman, M. Berry, R. Segev, and W. Bialek, Nature **440**, 1007 (2006).

[12] A. Braunstein and R. Zecchina, Phys. Rev. Lett. **96**, 030201 (2006).

[13] A. Braunstein, A. Pagnani, M. Weigt, and R. Zecchina, Journal of Statistical Mechanics: Theory and Experiment **2008**, P12001 (2008).

[14] Y. Roudi, J. A. Hertz, and E. Aurell, Front. Comput. Neurosci. **3** (2009).

[15] S. Cocco, S. Leibler, and R. Monasson, Proc Natl Acad Sci U S A **106**, 14058 (2009).

[16] V. Sessak and R. Monasson, J. Phys. A: Math. Theor. **42** (2009).

[17] M. Mézard and T. Mora, Journal of Physiology-Paris **103**, 107 (2009).

[18] E. Marinari and V. V. Kerrebroeck, J. Stat. Mech. (2010).

[19] S. Cocco and R. Monasson, Phys. Rev. Lett. **106**, 090601 (2011).

[20] F. Ricci-Tersenghi, J. Stat. Mech. (2012).

[21] H. Nguyen and J. Berg, J. Stat. Mech. (2012).

[22] H. Nguyen and J. Berg, Phys. Rev. Lett. **109** (2012).

[23] E. Aurell and M. Ekeberg, Phys. Rev. Lett. **108**, 090201 (2012).

[24] A. Hyvärinen, J. Karhunen, and E. Oja, *Independent Component Analysis*, Adaptive and Learning Systems for Signal Processing, Communications, and Control (John Wiley & Sons, 2001).

[25] J. Rissanen, *Information and Complexity in Statistical Modeling* (Springer, 2007).

[26] M. J. Wainwright and M. I. Jordan, Foundations and Trends in Machine Learning **1**, 1 (2008).

[27] P. Ravikumar, M. J. Wainwright, and J. D. Lafferty, Annals of Statistics **38**, 1287 (2010).

[28] H. Ackley, E. Hinton, and J. Sejnowski, Cognitive Science **9**, 147 (1985).

[29] G. Parisi, *Statistical Field Theory* (Addison Wesley, 1988).

[30] L. Peliti, *Statistical Mechanics in a Nutshell* (Princeton University Press, 2011), ISBN ISBN-13: 9780691145297.

[31] K. Fischer and J. Hertz, *Spin Glasses* (Cambridge University Press, 1993), ISBN 0521447771, 9780521447775.

[32] I. Pagani, K. Liolios, J. Jansson, I. Chen, T. Smirnova, B. Nosrat, and M. V.M., Nucleic Acids Res. **40**, D571 (2012).

[33] The Uniprot Consortium, Nucleic Acids Res. **40**, D71 (2012).

[34] H. Berman, G. Kleywegt, H. Nakamura, and J. Markley, Structure **20**, 391 (2012).

[35] U. Göbel, C. Sander, R. Schneider, and A. Valencia, Proteins: Struct., Funct., Genet. **18**, 309 (1994).

[36] S. W. Lockless and R. Ranganathan, Science **286**, 295 (1999).

[37] A. A. Fodor and R. W. Aldrich, Proteins: Structure, Function, and Bioinformatics **56**, 211 (2004).

[38] A. S. Lapedes, B. G. Giraud, L. Liu, and G. D. Stormo, Lecture Notes-Monograph Series: Statistics in Molecular Biology and Genetics **33**, pp. 236 (1999).

[39] J. Besag, The Statistician **24**, 179195 (1975).

[40] A. Schug, M. Weigt, J. Onuchic, T. Hwa, and H. Szurmant, Proc Natl Acad Sci USA **106**, 22124 (2009).

[41] A. E. Dago, A. Schug, A. Procaccini, J. A. Hoch, M. Weigt, , and H. Szurmant, Proc Natl Acad Sci USA **109**, 10148 (2012).

[42] T. A. Hopf, L. J. Colwell, R. Sheridan, B. Rost, C. Sander, and D. S. Marks, Cell **149**, 1607 (2012).

[43] G. Darmois, C.R. Acad. Sci. Paris **200**, 12651266 (1935), in French.

[44] E. Pitman and J. Wishart, Mathematical Proceedings of the Cambridge Philosophical Society **32**, 567579 (1936).

[45] B. Koopman, Transactions of the American Mathemati-

cal Society **39**, 399409 (1936).

[46] A. Kholodenko, Journal of Statistical Physics **58**, 35' (1990).

[47] T. Tanaka, Neural Computation **12**, 19511968 (2000).

[48] S. ichi Amari, S. Ikeda, and H. Shimokawa, *Information geometry and mean field approximation: the alpha projection approach* (MIT Press, 2001), chap. 16, pp 241–257, ISBN 0-262-15054-9.

[49] S. D. Dunn, L. M. Wahl, and G. B. Gloor, Bioinformatic **24**, 333 (2008).

[50] G. Hinton, Neural Computation **14**, 17711800 (2002).

[51] M. Gutmann and A. Hyvärinen, Journal of Machin Learning Research **13**, 307 (2012).

Appendix A: Further comparisons

Another way to visualize the comparative performanc of the two methods is *contact maps*, shown in Fig. 7. Th tendency observed in the scatter plots remains: NMFI-D has a larger portion of highly scored pairs in the neighbor zone (the middle stretch of the figures). Clusters o strongly interacting neighbors are caught by both algorithms, but PLM marks such sections using fewer pairs NMFI-DI displays somewhat loopy behavior in these regions; where PLM, for instance, identifies pairs of th type $(1, 2)$, $(2, 3)$, and $(3, 4)$, NMFI-DI tends to in addition include pairs like $(1, 3)$, $(1, 4)$, and $(2, 4)$, whicl could be argued to be somewhat redundant.

To get a sense of how false positives distribute acros the domains, we draw interactions into circles in Fig. 8. Among erroneously predicted contacts there is some tendency towards loopiness, especially for NMFI-DI; the blue lines tend to ‘bounce around’ in the circles. It hence seems that relatively few nodes are responsible for many of the false positives. We performed an explicit check of the data columns belonging to these ‘bad’ nodes, and we found that they often contained strongly biased data, i.e., had a few large $\mathbf{f}_i(k)$. In such cases, it seemed that NMFI-DI was more prone than PLM to report a (predicted) interaction.

Appendix B: Other scores for naive mean-field inversion

We investigated whether NMFI performance by using the APC term for the DI scoring and by using the CN score. In the second case we first switch the parameter constraints from (9) to (8) using (26). Mean TP rates using these modified scores are shown in fig. 9. We observe that APC in DI scoring increases TP rates slightly, while CN scoring can improve TP rates overall. We remark however that for for the second-highest ranked interaction ($p = 2$) NMFI with the original DI scoring (NMFI-DI) outperforms NMFI with CN scoring (NMFI-CN).

Motivated by the results of fig. 9, we decided to compare NMFI and PLM under the CN score. All figures in this paragraph show the best regularization for each method, unless otherwise stated. Figure 10 shows score vs. distance for all $|i-j| > 4$-pairs in all families. Unlike fig. 4a–b, the two plots now show very similar profiles. We note, however, that NMFI's CN scores trend two to three times larger than PLM's (the scales on the vertical axes are different). Perhaps this is an inherent feature for these methods, or simply a consequence of the different types of regularization types.

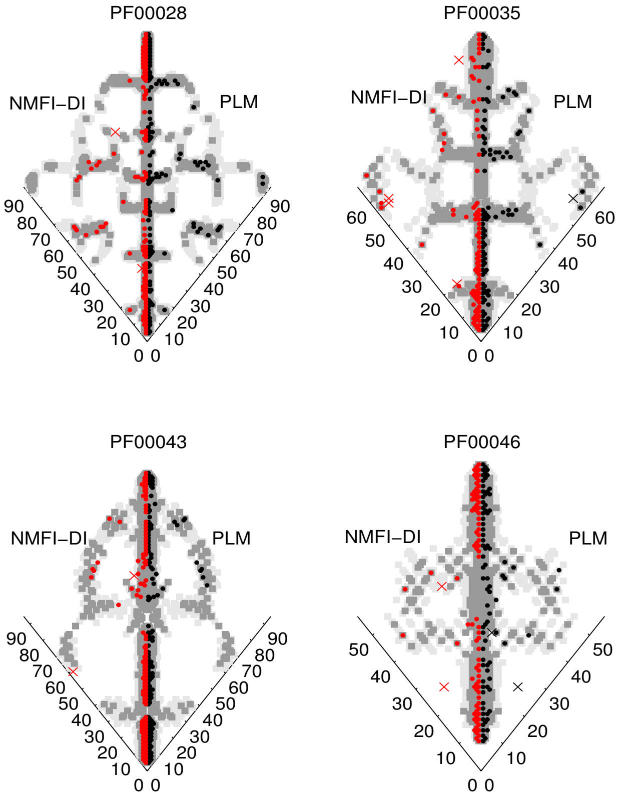


FIG. 7: Contact maps for PLM and NMFI-DI from four families. A pair (i, j)'s placement in the plots is found by matching positions i and j on the axes. Contacts are indicated by gray (dark for $d(i, j) < 5$Å and light for 5Å$\leq d(i, j) < 8.5$Å). True and false positives are represented by circles and crosses, respectively. Each figure shows the $1.5N$ strongest ranked pairs (including neighbors) for that family.

TP rates

Figure 11 shows the same situation as Fig. 3, but using CN to score NMFI. The three best regularization choices for NMFI-CN turned out the same as before, *i.e.* $\lambda = 1 \cdot B_{eff}$, $\lambda = 1.5 \cdot B_{eff}$ and $\lambda = 2.3 \cdot B_{eff}$, but the best out of these three was now $\lambda = 2.3 \cdot B_{eff}$ (instead of $\lambda = 1 \cdot B_{eff}$). Comparing Fig. 3 and Fig 11 one

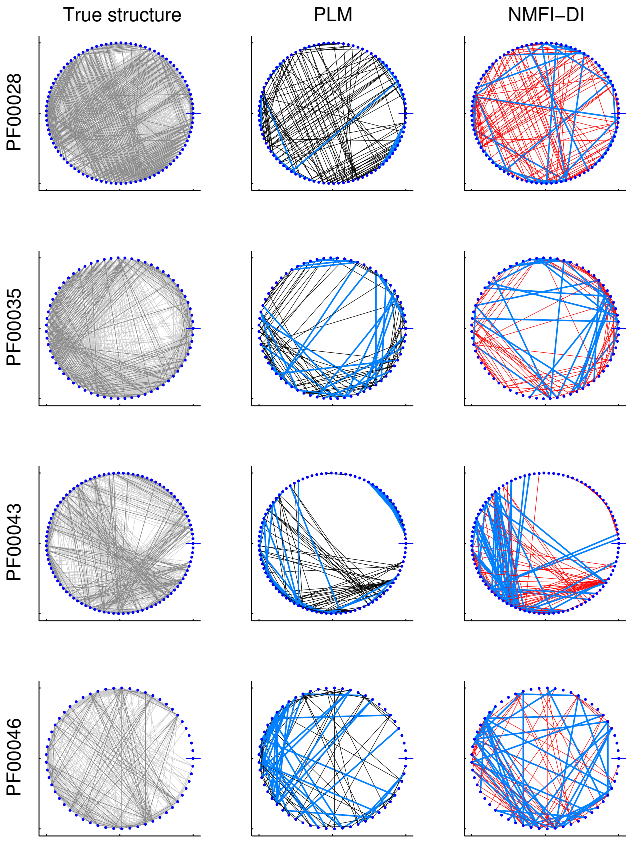


FIG. 8: Connections for four families overlaid on circles. Position '1' is indicated by a dash. The leftmost column shows contacts in the crystal structure (dark gray for $d(i,j) < 5$Å and light gray for 5Å$\leq d(i,j) < 8.5$Å). The other two columns show the top $1.5N$ strongest ranked $|i-j| > 4$-pairs for PLM and NMFI, with black/red for true positives and blue for false positives.

can see that the difference between the two methods is now smaller; for several families, the prediction quality is in fact about the same for both methods. Still, PLM maintains a somewhat higher TP rates overall.

Scatter plots

Figure 12 shows scatter plots for the same families as in Fig. 5 but using the CN scoring for NMFI. The points now lie more clearly on a line, from which we conclude that the bends in Fig. 5 was likely a consequence of differing scores. Yet, the trends seen in Fig. 5 remain: NMFI gives more attention to neighbor pairs than does PLM.

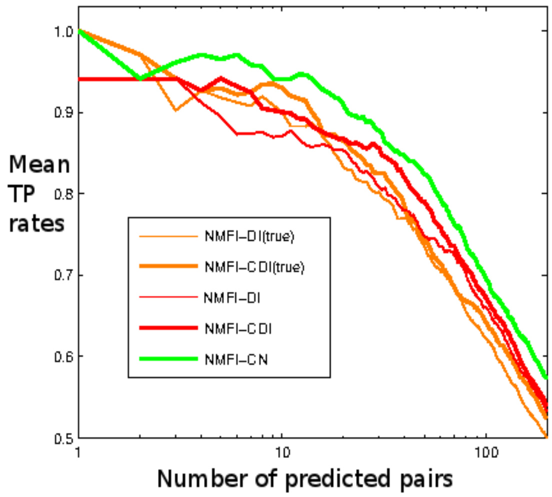


FIG. 9: Mean TP rates, using pairs with $|i-j| > 4$, for NMFI with old scores DI and DI(true), new APC scores CDI and CDI(true), and the norm score CN. Each curve corresponds to the best λ for that particular score.

Contact maps

In Fig. 13 we recreate the contact maps of fig. 7 with NMFI-CN in place of NMFI-DI and find that the plots are more symmetric. As expected, asymmetry is seen primarily for small $|i-j|$; NMFI tends to crowd these regions with lots of loops.

Gap-gap interactions

To investigate why NMFI assembles so many top-scored pairs in certain neighbor regions, we performed an explicit check of the associated MSA columns. A relevant regularity was observed: when gaps appear in a sequence, they tend to do so in long strands. The picture can be illustrated by the following hypothetical MSA (in our implementation, the gap state is 1):

```
... 6 5 9 7 2 6 8 7 4 4 2 2 ...
... 1 1 1 1 1 1 1 1 1 1 2 8 ...
... 6 5 2 7 2 3 8 9 5 4 2 3 ...
... 3 7 4 7 2 6 8 7 9 4 2 3 ...
... 3 7 4 7 2 3 8 8 9 4 2 9 ...
... 1 1 1 1 1 1 1 4 5 4 2 9 ...
... 8 5 9 7 2 9 8 7 4 4 2 4 ...
... 1 1 1 1 1 1 1 1 1 1 2 4 ...
```

We recall that gaps ("1" states) are necessary for satisfactory alignment of the sequences in a family and that in our procedure we treat gaps just another amino acid, with its associated interaction parameters. We then make the obvious observation that independent samples from a Potts model will only contain long subsequences

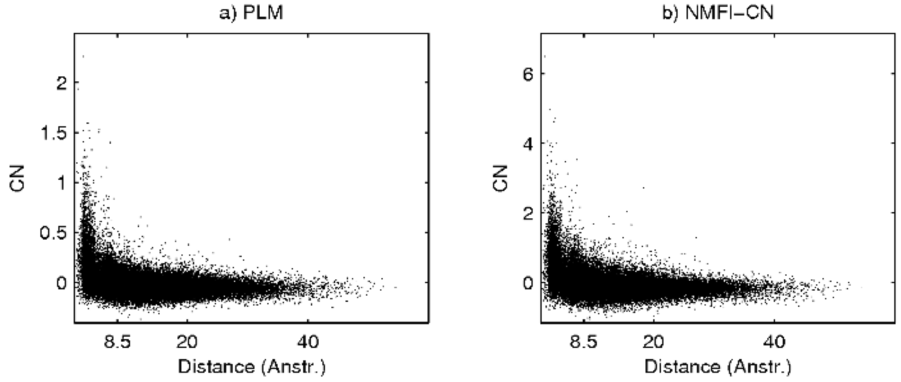


FIG. 10: Score plotted against distance for all $|i-j| > 4$-pairs in all 17 families.

of the same state with low probability. In other words, the model to which we fit the data cannot describe long stretches of "1" states, which is a feature of the data. It is hence quite conceivable that the two methods handle this discrepancy between data and models differently since we do expect this gap effect to generate large $\mathbf{J}_{ij}(1,1)$ for at least some pairs with small $|i-j|$.

Figure 14 shows scatter plots for all coupling parameters $\mathbf{J}_{ij}(k,l)$ in PF00014, which has a modest amount of gap sections, and in PF00043, which has relatively many. As outlines above, the $\mathbf{J}_{ij}(1,1)$-parameters are among the largest in magnitude, especially for PF00043. We also note that the red dots steer to the right; NMFI clearly reacts harder to the gap-gap interactions than PLM.

Jones *et al.* (2012) disregarded contributions from gaps in their scoring by simply skipping the gap state when doing their norm summations. We tried this but found no significant improvement for either method. The change seemed to affect only pairs with small $|i-j|$ (which is reasonable), and our TP rates are based on pairs with $|i-j| > 4$. If gap interactions are indeed responsible for reduced prediction qualities, removing their input during scoring is just a band-aid type solution. An better way would be to suppress them already in the parameter estimation step. That way, all interplay would have to be accounted for without them. Whether or not there are ways to effectively handle the inference problem in PSP ignording gaps or treating them differently is an issue which goes beyond the scope of this work.

We also investigated whether the gap effect depends on the sequence similarity re-weighting factor x, which up to here was chosen $x = 0.9$. Perhaps the gap effect can be dampened by stricter definition of sequence uniqueness? In Fig. 15 we show another set of TP rates, but now for $x = 0.75$. We also include results for NMFI run on alignment files from which all sequences with more than 20% gaps have been removed. The best regularization choice for each method turned out the same as in fig. 11: $\lambda = 2.3 \cdot B_{eff}$ for NMFI and $\lambda_J = 0.01$ for PLM. Overall, PLM keeps the same advantage over NMFI it had in Fig. 11. Removing gappy sequences seems to trim down more TP rates than it raises, probably since useful information in the non-gappy parts is discarded unnecessarily.

Appendix C: Extension to 28 protein families

To sample a larger set of families, we conducted an additional survey of 28 families, now covering lengths across the wider range of 50-400. The list is given in Table C. We here kept the re-weighting level at $x = 0.8$ as in [2], while the TP rates were again calculated using the cutoff 8.5Å. The pseudo-count strength for NMFI was varied in the same interval as in the main text. We did not try to optimize the regularization parameters for this trial, but merely used $\lambda_J = 0.01$ as determined for the smaller family in the main text.

Figure 16 shows qualitatively the same behaviour as in the smaller set of families: TP rates increase partly from changing the from the DI score to CN score, and partly from changing from NMFI to PLM. Our positive results thus do *not* seem to be particular to short-length families.

Apart from the average TP rate for each value of p (p'th strongest predicted interaction) one can also evaluate performance by different criteria. In this large family we investigated the distribution of values of p such that the TP rate in a family is one. Fig. 17 shows the histograms of the number of families for which the top p predictions are correct, clearly showing that the difference between PLM and NMFI (using the two scores) primarily occurs at the high end. The difference in average performance between PLM and NMFI at least partially stems from PLM getting more strongest contact predictions with 100% accuracy.

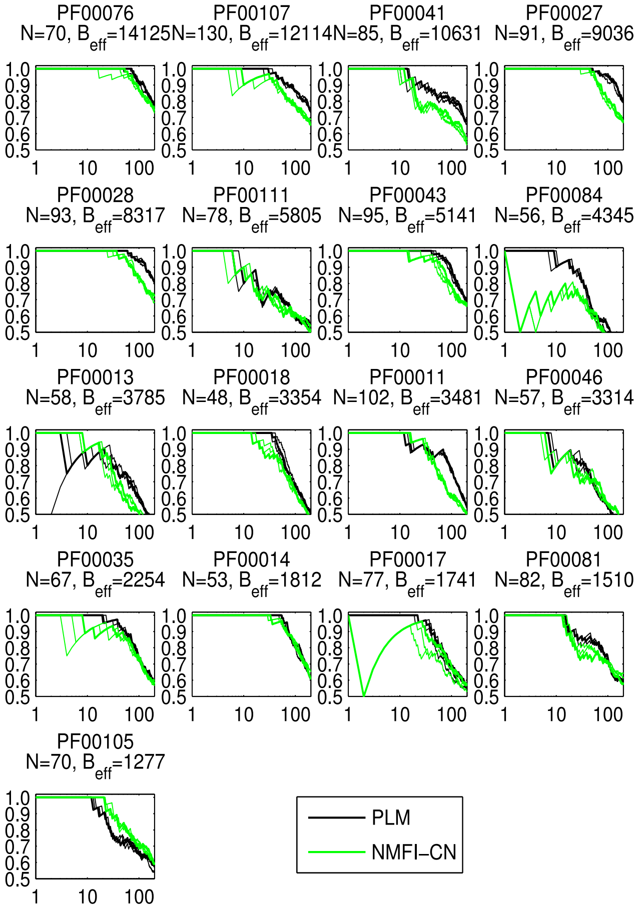


FIG. 11: Contact-detection results for all the families in our study (sorted by B_{eff}), now with the CN score for NMFI. Y-axes are TP rates and x-axes are the number of predicted contacts p, based on pairs with $|i-j| > 4$. The three curves for each method are the three regularization levels yielding highest TP rates across all families. The thickened curve highlights the best one out of these three ($\lambda = 2.3 \cdot B_{eff}$ for NMFI-CN and $\lambda_J = 0.01$ for PLM).

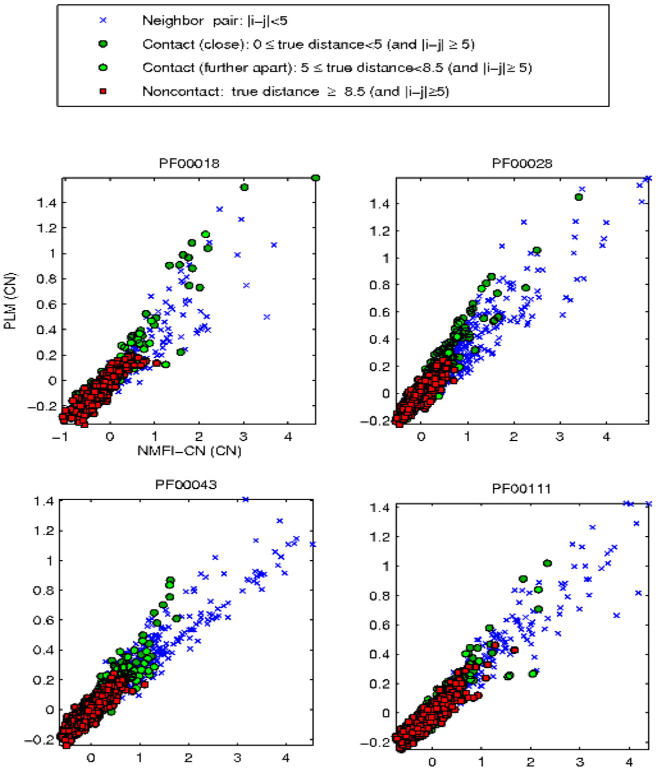


FIG. 12: Scatter plots of interaction scores for PLM and NMFI-CN from four families. For all plots, the axes are as indicated by the top left one. The distance unit in the top box is Å.

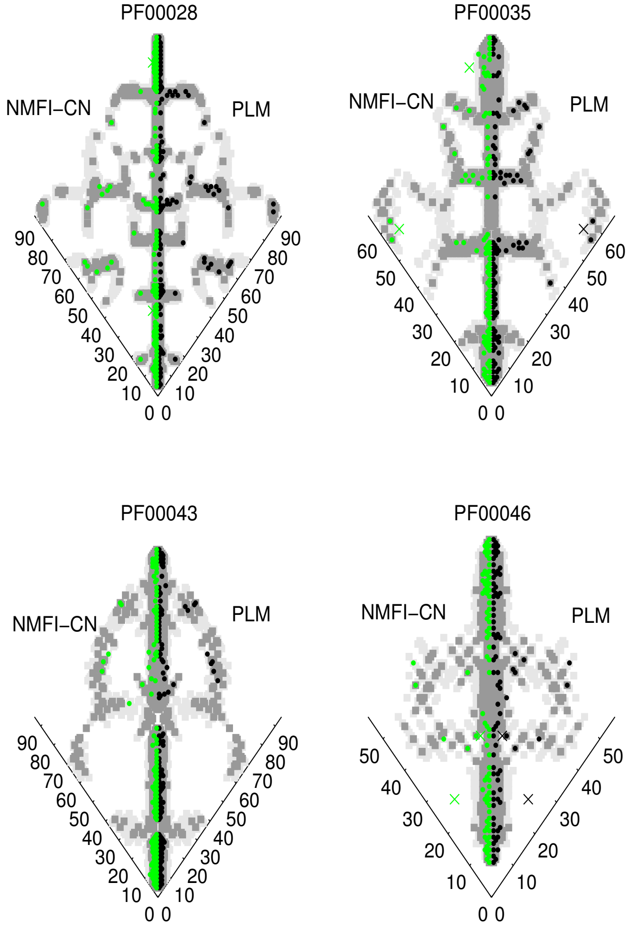


FIG. 13: Contact maps for PLM and NMFI-CN from four families. A pair (i, j)'s placement in the plots is found by matching positions i and j on the axes. Contacts are indicated by gray (dark for $d(i, j) < 5$Å and light for 5Å$\leq d(i, j) <$ 8.5Å). True and false positives are represented by circles and crosses, respectively. Each figure shows the $1.5N$ strongest ranked pairs (including neighbors) for that family.

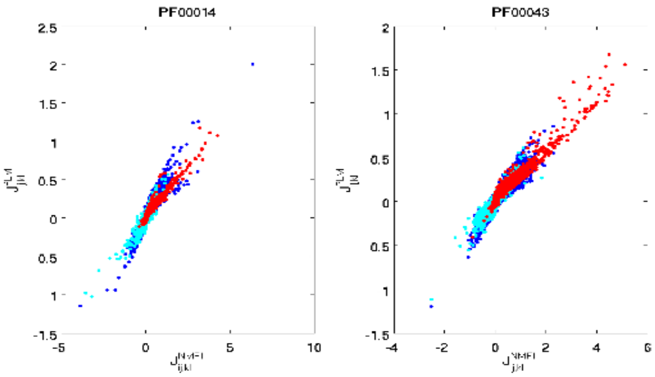


FIG. 14: Scatter plots of estimated $J_{ij,kl} = \mathbf{J}_{ij}(k,l)$ from PF00014 and PF00043. Red dots are ‘gap–gap’ interactions ($k = l = 1$), turquoise dots are ‘gap–amino-acid’ interactions ($k = 1$ and $l \neq 1$, or $k \neq 1$ and $l = 1$), and blue dots are ‘amino-acid–amino-acid’ interactions ($k \neq 1$ and $l \neq 1$).

ID	N	B	$B_{eff}(80\%)$
PF00006	215	10765	640.68
PF00011	102	5024	2725.01
PF00013	58	6059	2529.25
PF00014	53	2393	1478.23
PF00017	77	2732	1311.75
PF00018	48	5073	334.68
PF00025	175	2946	995.75
PF00026	314	3851	2074.68
PF00027	91	12129	7631.12
PF00028	93	12628	6322.61
PF00032	102	14994	684.47
PF00035	67	3093	1826.30
PF00041	85	15551	8691.38
PF00043	95	6818	4051.93
PF00044	151	6206	1422.27
PF00046	57	7372	1760.85
PF00056	142	4185	1119.53
PF00059	108	5293	3258.25
PF00071	161	10779	3793.01
PF00073	171	9524	487.07
PF00076	70	21125	10112.69
PF00081	82	3229	890.25
PF00084	56	5831	3453.40
PF00085	104	10569	
PF00091	216	8656	916.98
PF00092	179	3936	1785.97
PF00105	70	2549	816.12
PF00108	264	6839	2688.27

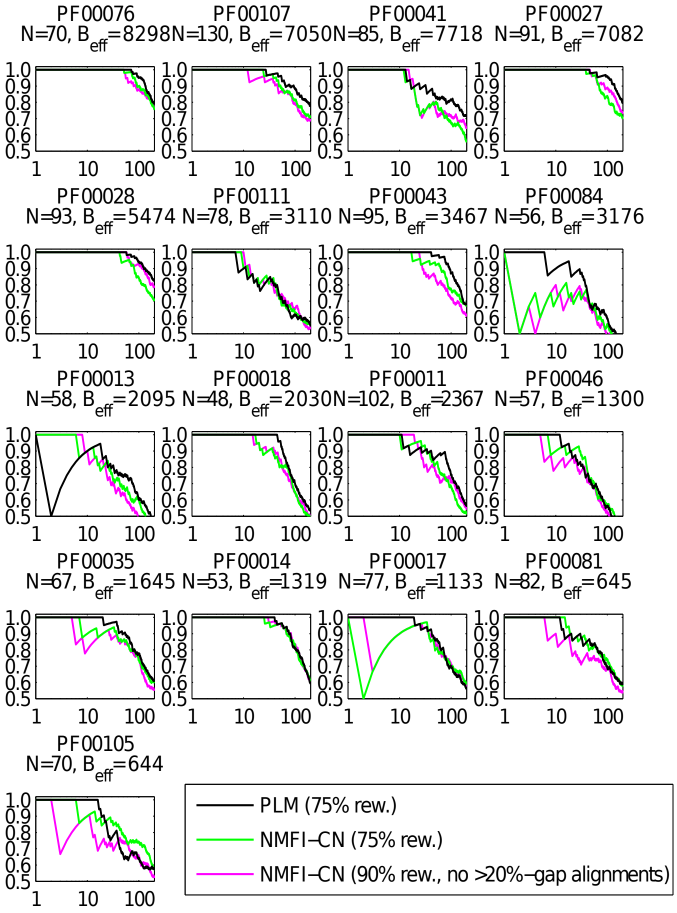


FIG. 15: Contact-detection results for all the families in our study. Y-axes are TP rates and x-axes are the number of predicted contacts p, based on pairs with $|i-j| > 4$. The black and green curves are for re-weighting margin $x = 0.75$, and the purple curve is for re-weighting margin $x = 0.9$ after deletion of all sequences with more than 20% gaps. The curve for each method corresponds to the regularization level yielding highest TP rates across all families ($\lambda = 2.3 \cdot B_{eff}$ for both NMFI-CN and $\lambda_J = 0.01$ for PLM).

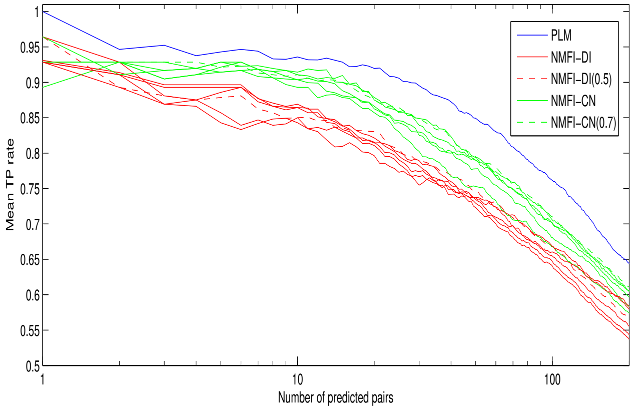


FIG. 16: Mean TP rates over the larger set of 28 families for PLM with $\lambda_J = 0.01$ (blue), and varying regularization values for NMFI-CN (green) and NMFI-DI (red).

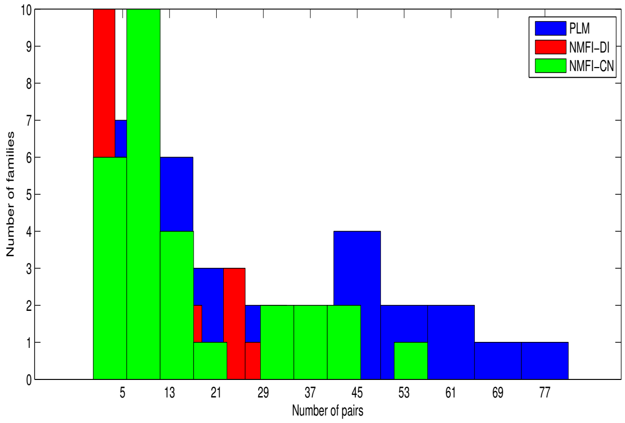


FIG. 17: Distribution of 'perfect' accuracy for the three methods. The x-axis shows the number of top-ranked pairs for which the TP rates stays at one, and the y-axis shows the number of families.