REVIEW

# Inferring Pairwise Interactions from Biological Data Using Maximum-Entropy Probability Models

**Richard R. Stein**[1]*, **Debora S. Marks**[2], **Chris Sander**[1]*

[1] Computational Biology Program, Sloan Kettering Institute, Memorial Sloan Kettering Cancer Center, New York, New York, United States of America, [2] Department of Systems Biology, Harvard Medical School, Boston, Massachusetts, United States of America

* maxent@sanderlab.org (RRS); maxent@sanderlab.org (CS)

OPEN ACCESS

**Citation:** Stein RR, Marks DS, Sander C (2015) Inferring Pairwise Interactions from Biological Data Using Maximum-Entropy Probability Models. PLoS Comput Biol 11(7): e1004182. doi:10.1371/journal.pcbi.1004182

**Editor:** Shi-Jie Chen, University of Missouri, UNITED STATES

**Published:** July 30, 2015

**Funding:** This work was supported by NIH awards R01 GM106303 (DSM, CS, and RRS) and P41 GM103504 (CS). The funders had no role in the preparation of the manuscript.

**Competing Interests:** The authors have declared that no competing interests exist.

## Abstract

Maximum entropy-based inference methods have been successfully used to infer direct interactions from biological datasets such as gene expression data or sequence ensembles. Here, we review undirected pairwise maximum-entropy probability models in two categories of data types, those with continuous and categorical random variables. As a concrete example, we present recently developed inference methods from the field of protein contact prediction and show that a basic set of assumptions leads to similar solution strategies for inferring the model parameters in both variable types. These parameters reflect interactive couplings between observables, which can be used to predict global properties of the biological system. Such methods are applicable to the important problems of protein 3-D structure prediction and association of gene–gene networks, and they enable potential applications to the analysis of gene alteration patterns and to protein design.

## Introduction

Modern high-throughput techniques allow for the quantitative analysis of various components of the cell. This ability opens the door to analyzing and understanding complex interaction patterns of cellular regulation, organization, and evolution. In the last few years, **undirected pairwise maximum-entropy probability models** have been introduced to analyze biological data and have performed well, disentangling **direct interactions** from artifacts introduced by intermediates or spurious coupling effects. Their performance has been studied for diverse problems, such as gene network inference [1,2], analysis of neural populations [3,4], protein contact prediction [5–8], analysis of a text corpus [9], modeling of animal flocks [10], and prediction of multidrug effects [11]. Statistical inference methods using partial correlations in the context of graphical Gaussian models (GGMs) have led to similar results and provide a more intuitive understanding of direct versus indirect interactions by employing the concept of conditional independence [12,13].

Our goal here is to derive a unified framework for pairwise maximum-entropy probability models for continuous and categorical variables and to discuss some of the recent inference approaches presented in the field of protein contact prediction. The structure of the manuscript is as follows: (1) introduction and statement of the problem, (2) deriving the probabilistic model, (3) inference of interactions, (4) scoring functions for the pairwise interaction strengths, and (5) discussion of results, improvements and applications.

Better knowledge of these methods, along with links to existing implementations in terms of software packages, may be helpful to improve the quality of biological data analysis compared to standard correlation-based methods and increase our ability to make predictions of interactions that define the properties of a biological system. In the following, we highlight the power of inference methods based on the maximum-entropy assumption using two examples of biological problems: inferring networks from gene expression data and residue contacts in proteins from multiple sequence alignments. We compare solutions obtained using (1) correlation-based inference and (2) inference based on pairwise maximum-entropy probability models (or their incarnation in the continuous case, the multivariate Gaussian distribution).

## Gene association networks

Pairwise associations between genes and proteins can be determined by a variety of data types, such as gene expression or protein abundance. Association between entities in these data types are commonly estimated by the sample **Pearson correlation** coefficient computed for each pair of variables $x_i$ and $x_j$ from the set of random variables $x_1$,. . ., $x_L$. In particular, for $M$ given samples in $L$ measured variables, $\mathbf{x}^1 = (x_1^1, \ldots, x_L^1)^{\mathrm{T}}, \ldots, \mathbf{x}^M = (x_1^M, \ldots, x_L^M)^{\mathrm{T}} \in \mathbb{R}^L$, it is defined as,

$$r_{ij} := \frac{\hat{C}_{ij}}{\sqrt{\hat{C}_{ii}\hat{C}_{jj}}},$$

where $\hat{C}_{ij} := \frac{1}{M}\sum_{m=1}^{M}(x_i^m - \overline{x_i})(x_j^m - \overline{x_j})$ denotes the $(i, j)$-element of the empirical covariance matrix $\hat{C} = (\hat{C}_{ij})_{i,j=1,\ldots,L}$. The sample mean operator $\overline{\cdot}$ provides the empirical mean from the measured data and is defined as $\overline{x_i} := \frac{1}{M}\sum_{m=1}^{M} x_i^m$. A simple way to characterize dependencies in data is to classify two variables as being dependent if the absolute value of their correlation coefficient is above a certain threshold (and independent otherwise) and then use those pairs to draw a so-called relevance network [14]. However, the Pearson correlation is a misleading measure for direct dependence as it only reflects the association between two variables while ignoring the influence of the remaining ones. Therefore, the relevance network approach is not suitable to deduce direct interactions from a dataset [15–18]. The **partial correlation** between two variables removes the variational effect due to the influence of the remaining variables (Cramér [19], p. 306). To illustrate this, let's take a simplified example with three random variables $x_{\mathrm{A}}$, $x_{\mathrm{B}}$, $x_{\mathrm{C}}$. Without loss of generality, we can scale each of these variables to zero-mean and unit-standard deviation by $x_i \mapsto (x_i - \overline{x_i}) \Big/ \sqrt{\hat{C}_{ii}}$, which simplifies the correlation coefficient to $r_{ij} \equiv \overline{x_i x_j}$. The sample partial correlation coefficient of a three-variable system between $x_{\mathrm{A}}$ and $x_{\mathrm{B}}$ *given* $x_{\mathrm{C}}$ is then defined as [19,20]

$$r_{\mathrm{AB\cdot C}} = \frac{r_{\mathrm{AB}} - r_{\mathrm{BC}} r_{\mathrm{AC}}}{\sqrt{1 - r_{\mathrm{AC}}^2}\sqrt{1 - r_{\mathrm{BC}}^2}} \equiv -\frac{(\hat{C}^{-1})_{\mathrm{AB}}}{\sqrt{(\hat{C}^{-1})_{\mathrm{AA}}(\hat{C}^{-1})_{\mathrm{BB}}}}.$$

The latter equivalence by Cramer's rule holds if the empirical covariance matrix, $\hat{C} = (\hat{C}_{ij})_{i,j\in\{\mathrm{A,B,C}\}}$, is invertible. Krumsiek et al. [21] studied the Pearson correlations and partial correlations in data generated by an *in silico* reaction system consisting of three components A, B, C with reactions between A and B, and B and C (Fig 1A). A graphical comparison of

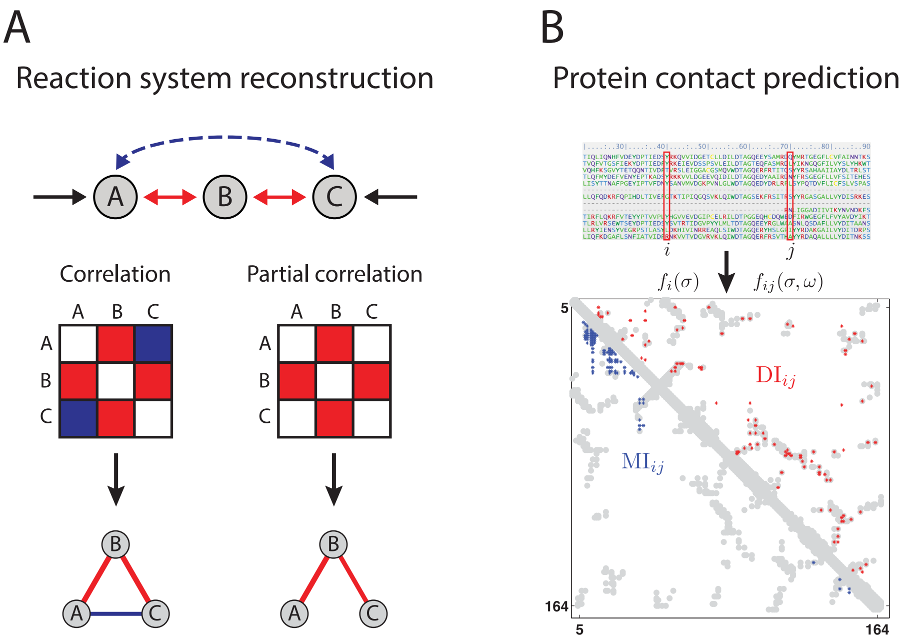

**Fig 1. Reaction system reconstruction and protein contact prediction.** Association results of correlation-based and maximum-entropy methods on biological data from an *in silico* reaction system (A) and protein contacts (B). (A) Analysis by Pearson's correlation yields interactions associating all three compounds A, B, and C, in contrast to the partial correlation approach which omits the "false" link between A and C. (Fig 1A based on [21].) (B) Protein contact prediction for the human RAS protein using the correlation-based mutual information, MI, and the maximum-entropy based direct information, DI, (blue and red, respectively). The 150 highest scoring contacts from both methods are plotted on the protein contacts from experimentally determined structure in gray. (Fig 1B based on [6].)

doi:10.1371/journal.pcbi.1004182.g001

Pearson's correlations, $r_{\mathrm{AB}}$, $r_{\mathrm{AC}}$ $r_{\mathrm{BC}}$, versus the corresponding partial correlations, $r_{\mathrm{AB \cdot C}}$, $r_{\mathrm{AC \cdot B}}$, $r_{\mathrm{BC \cdot A}}$, shows that variables A and C appear to be correlated when using Pearson's correlation as a dependency measure since both are highly correlated with variable B, which results in a false inferred reaction $r_{\mathrm{AC}}$. The strength of the incorrectly inferred interaction can be numerically large and therefore particularly misleading if there are multiple intermediate variables B [22]. The partial correlation analysis removes the effect of the mediating variable(s) B and correctly recovers the underlying interaction structure. This is always true for variables following a multivariate Gaussian distribution, but also seems to work empirically on realistic systems as Krumsiek et al. [21] have shown for more complex reaction structures than the example presented here.

## Protein contact prediction

The idea that protein contacts can be extracted from the evolutionary family record was formulated and tested some time ago [23–26]. The principle used here is that slightly deleterious mutations are compensated during evolution by mutations of residues in contact in order to maintain the function and, by implication, the shape of the protein. Protein residues that are close in space in the folded protein are often mutated in a correlated manner. The main problem here is that one has to disentangle the directly co-evolving residues and remove transitive correlations from the large number of other co-variations in protein sequences that arise due to statistical noise or phylogenetic sampling bias in the sequence family. Interactions not internal to the protein are, for example, evolutionary constraints on residues involved in oligomerization, protein–protein, protein–substrate interactions [6,27,28]. In particular, the empirical single-site and pair frequency counts in residue $i$ and in residues $i$ and $j$ for elements $\sigma$, $\omega$ of the 20-element amino acid alphabet plus gap, $f_i(\sigma)$ and $f_{ij}(\sigma, \omega)$, are extracted from a representative multiple sequence alignment under applied reweighting to account for biases due to undersampling. Correlated evolution in these positions was analyzed, e.g., by [29], by using the **mutual information** between residue $i$ and $j$,

$$\mathrm{MI}_{ij} = \sum_{\sigma,\omega} f_{ij}(\sigma,\omega) \ln\left(\frac{f_{ij}(\sigma,\omega)}{f_i(\sigma) f_j(\omega)}\right).$$

Although results did show promise, an important improvement was made years later by using a maximum-entropy approach on the same setup [5–7,30]. In this framework, the **direct information** of residue $i$ and $j$ was introduced by replacing $f_{ij}$ in the mutual information by $P_{ij}^{\mathrm{dir}}$,

$$\mathrm{DI}_{ij} = \sum_{\sigma,\omega} P_{ij}^{\mathrm{dir}}(\sigma,\omega) \ln\left(\frac{P_{ij}^{\mathrm{dir}}(\sigma,\omega)}{f_i(\sigma) f_j(\omega)}\right), \tag{1}$$

where $P_{ij}^{\mathrm{dir}}(\sigma,\omega) = \frac{1}{Z_{ij}} \exp(e_{ij}(\sigma,\omega) + \tilde{h}_i(\sigma) + \tilde{h}_j(\omega))$ and $\tilde{h}_i(\sigma), \tilde{h}_j(\omega)$ and $Z_{ij}$ are chosen such that $P_{ij}^{\mathrm{dir}}$, which is based on a pairwise probability model of an amino acid sequence compatible with the iso-structural sequence family, is consistent with the single-site frequency counts. In an approximative solution, [6,7] determined the contact strength between the amino acids $\sigma$ and $\omega$ in position $i$ and $j$, respectively, by

$$e_{ij}(\sigma,\omega) \simeq -(C^{-1}(\sigma,\omega))_{ij}. \tag{2}$$

Here, $(C^{-1}(\sigma,\omega))_{ij}$ denotes the inverse element corresponding to $C_{ij}(\sigma,\omega) \equiv f_{ij}(\sigma,\omega) - f_i(\sigma) f_j(\omega)$ for amino acids $\sigma$, $\omega$ from a subset of 20 out of the 21 different states (the so-called gauge fixing, see below). The comparison of contact prediction results based on MI- and DI-score for the RAS human protein on top of the actual crystal structure shows a much more accurate prediction result when using the direct information instead of the mutual information (Fig 1B).

The next section lays the foundation to deriving maximum-entropy models for the two data types: continuous, as used in the first example, and categorical, as used in the second one. Subsequently, we will present inference techniques to solve for their interaction parameters.

## Deriving the Probabilistic Model

Ideally, one would like to use a probabilistic model that is, on the one hand, able to capture all orders of directed interactions of all observables at play and, on the other hand, correctly reproduces the observed and to-be-predicted frequencies. However, this would require a prohibitively large number of observed data points. For this reason, we restrict ourselves to probabilistic models with terms up to second order, which we derive for continuous, real-valued variables, and extend this framework to models with categorical variables that are suitable, for example, to treat sequence information in the next section.

### Model formulation for continuous random variables

We model the occurrence of sets of events in a particular biological system by a multivariate probability distribution $P(\mathbf{x})$ of $L$ random variables $\mathbf{x} = (x_1,\ldots,x_L)^{\mathrm{T}} \in \mathbb{R}^L$ that is, on the one hand, consistent with the mean and covariance obtained from $M$ observed data values $\mathbf{x}^1,\ldots,\mathbf{x}^M$ and, on the other hand, maximizing the information entropy, $S$, to obtain the simplest possible probability model consistent with the data. At this point, each of the data's variables $x_i$ is continuously distributed on real values. In a biological example, these data originate from gene expression studies and each variable $x_i$ corresponds to the normalized mRNA level of a gene measured in $M$ samples. As an example, a recent pan-cancer study of The Cancer Genome Atlas (TCGA) provided mRNA levels from $M = 3{,}299$ patient tumor samples from 12 cancer types [31]. The problem can be large, e.g., in the case of a gene–gene association study one has $L \approx 20{,}000$ human genes.

The first constraint on the unknown probability distribution, $P: \mathbb{R}^L \to \mathbb{R}_{\geq 0}$ is that its integral normalizes to 1,

$$\int_{\mathbf{x}} P(\mathbf{x})\, \mathrm{d}\mathbf{x} = 1, \tag{3}$$

which is a natural requirement on any probability distribution. Additionally, the first moment of variable $x_i$ is supposed to match the value of the corresponding sample mean over $M$ measurements in each $i = 1,\ldots, L$,

$$\langle x_i \rangle = \int_{\mathbf{x}} P(\mathbf{x}) x_i \, \mathrm{d}\mathbf{x} = \frac{1}{M} \sum_{m=1}^{M} x_i^m = \overline{x_i}, \tag{4}$$

where we define the $n$-th moment of the random variable $x_i$ distributed by the multivariate probability distribution $P$ as $\langle x_i^n \rangle := \int_{\mathbf{x}} P(\mathbf{x}) x_i^n \, \mathrm{d}\mathbf{x}$. Analogously, the second moment of the variables $x_i$ and $x_j$ and its corresponding empirical expectation is supposed to be equal,

$$\langle x_i x_j \rangle = \int_{\mathbf{x}} P(\mathbf{x}) x_i x_j \, \mathrm{d}\mathbf{x} = \frac{1}{M} \sum_{m=1}^{M} x_i^m x_j^m = \overline{x_i x_j} \tag{5}$$

for $i, j = 1,\ldots, L$. Taken together, Eqs 4 and 5 constrain the distribution's covariance matrix to be coherent to the empirical covariance matrix. Finally, the probability distribution should maximize the information entropy,

$$\text{maximize } S = -\int_{\mathbf{x}} P(\mathbf{x}) \ln P(\mathbf{x}) \, \mathrm{d}\mathbf{x} \tag{6}$$

with the natural logarithm ln. A well-known analytical strategy to find functional extrema under equality constraints is the **method of Lagrange multipliers** [32], which converts a constrained optimization problem into an unconstrained one by means of the Lagrangian $\mathcal{L}$. In our case, the probability distribution maximizing the entropy (Eq 6) subject to Eqs 3–5 is found as the stationary point of the Lagrangian $\mathcal{L} = \mathcal{L}(P(\mathbf{x}); \alpha, \boldsymbol{\beta}, \boldsymbol{\gamma})$ [33,34],

$$\mathcal{L} = S + \alpha(\langle 1 \rangle - 1) + \sum_{i=1}^{L} \beta_i (\langle x_i \rangle - \overline{x_i}) + \sum_{i,j=1}^{L} \gamma_{ij} (\langle x_i x_j \rangle - \overline{x_i x_j}). \tag{7}$$

The real-valued Lagrange multipliers $\alpha$, $\boldsymbol{\beta} = (\beta_i)_{i = 1,\ldots, L}$ and $\boldsymbol{\gamma} = (\gamma_{ij})_{i,j = 1,\ldots, L}$ correspond to the constraints Eqs 3, 4, and 5, respectively. The maximizing probability distribution is then found by setting the functional derivative of $\mathcal{L}$ with respect to the unknown density $P(\mathbf{x})$ to zero [33,35],

$$\frac{\delta \mathcal{L}}{\delta P(\mathbf{x})} = 0 \quad \Rightarrow \quad -\ln P(\mathbf{x}) - 1 + \alpha + \sum_{i=1}^{L} \beta_i x_i + \sum_{i,j=1}^{L} \gamma_{ij} x_i x_j = 0.$$

Its solution is the **pairwise maximum-entropy probability distribution**,

$$P(\mathbf{x};\boldsymbol{\beta},\boldsymbol{\gamma}) = \exp\left(-1+\alpha+\sum_{i=1}^{L}\beta_i x_i + \sum_{i,j=1}^{L}\gamma_{ij}x_i x_j\right) = \frac{1}{Z}\mathrm{e}^{-\mathcal{H}(\mathbf{x};\boldsymbol{\beta},\boldsymbol{\gamma})} \tag{8}$$

which is contained in the family of exponential probability distributions and assigns a non-negative probability to any system configuration $\mathbf{x} = (x_1,\ldots,x_L)^{\mathrm{T}} \in \mathbb{R}^L$. For the second identity, we introduced the **partition function** as normalization constant,

$$Z(\boldsymbol{\beta},\boldsymbol{\gamma}) := \int_{\mathbf{x}} \exp\left(\sum_{i=1}^{L}\beta_i x_i + \sum_{i,j=1}^{L}\gamma_{ij}x_i x_j\right)\mathrm{d}\mathbf{x} \equiv \exp(1-\alpha)$$

with the Hamiltonian, $\mathcal{H}(\mathbf{x}) := -\sum_{i=1}^{L}\beta_i x_i - \sum_{i,j=1}^{L}\gamma_{ij}x_i x_j$. It can be shown by means of the information inequality that Eq 8 is the unique maximum-entropy distribution satisfying the constraints Eqs 3–5 (Cover and Thomas [35], p. 410). Note that $\alpha$ is fully determined for given $\boldsymbol{\beta} = (\beta_i)$ and $\boldsymbol{\gamma} = (\gamma_{ij})$ by the normalization constraint Eq 3 and is therefore not a free parameter. The right-hand representation of Eq 8 is also referred to as **Boltzmann distribution**. The matrix of Lagrange multipliers $\boldsymbol{\gamma} = (\gamma_{ij})$ has to have full rank in order to ensure a unique parametrization of $P(\mathbf{x})$, otherwise, one can eliminate dependent constraints [33,36]. In addition, for the integrals in Eqs 3–6 to converge with respect to $L$-dimensional Lebesgue measure, we require $\boldsymbol{\gamma}$ to be negative definite, i.e., all of its eigenvalues to be negative or $\sum_{i,j}\gamma_{ij}x_i x_j = \mathbf{x}^{\mathrm{T}}\boldsymbol{\gamma}\mathbf{x} < 0$ for $\mathbf{x} \neq \mathbf{0}$.

## Concept of entropy maximization

Shannon states in his seminal work that information and (information) entropy are linked: the more information is encoded in the system, the lower its entropy [37]. Jaynes introduced the entropy maximization principle, which selects for the probability distribution that is (1) in agreement with the measured constraints and (2) contains the least information about the probability distribution [38–40]. In particular, any unnecessary information would lower the entropy and, thus, introduce biases and allow overfitting. As demonstrated in the section above, the assumption of entropy maximization under first and second moment constraints results in an exponential model or Markov random field (in log-linear form) and many of the properties shown here can be generalized to this model class [41]. On the other hand, there is some analogy of entropy as introduced by Shannon to the thermodynamic notion of entropy. Here, the Second law of Thermodynamics states that each isolated system monotonically evolves in time towards a state of maximum entropy, the equilibrium. A thorough discussion of this analogy and its limitation in non-equilibrium systems is beyond the scope of this review, but can be found in [42,43]. Here, we exclusively use the notion entropy maximization as the principle of minimal information content in the probability model consistent with the data.

## Categorical random variables

In the following section, we derive the pairwise maximum-entropy probability distribution on categorical variables. For jointly distributed categorical variables $\mathbf{x} = (x_1,\ldots,x_L)^{\mathrm{T}} \in \Omega^L$, each variable $x_i$ is defined on the finite set $\Omega = \{\sigma_1,\ldots,\sigma_q\}$ consisting of $q$ elements. In the concrete example of modeling protein co-evolution, this set contains the 20 amino acids represented by a 20-letter alphabet from $A$ standing for Alanine to $Y$ for Tyrosine plus one gap element, then $\Omega = \{A, C, D, E, F, G, H, I, K, L, M, N, P, Q, R, S, T, V, W, Y, –\}$ and $q = 21$. Our goal is to extract co-evolving residue pairs from the evolutionary record of a given protein family. As input data,

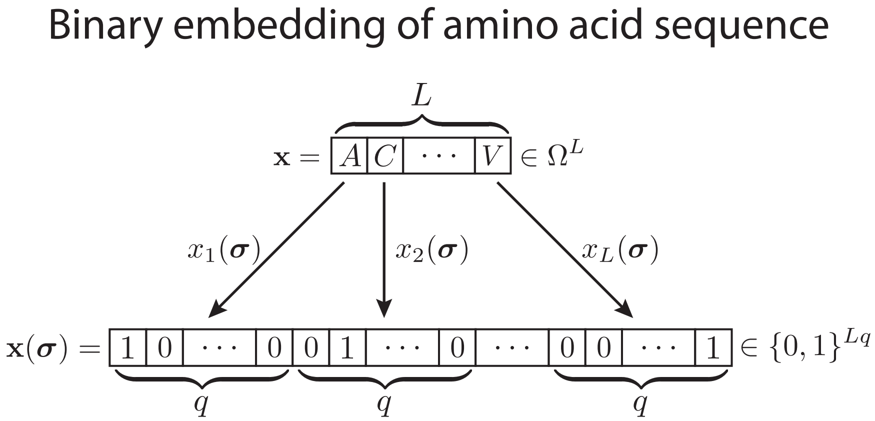
Binary embedding of amino acid sequence
L
$\mathbf{x} = \boxed{A|C|\cdots|V} \in \Omega^L$
$x_1(\boldsymbol{\sigma})$
$x_2(\boldsymbol{\sigma})$
$x_L(\boldsymbol{\sigma})$
$\mathbf{x}(\boldsymbol{\sigma}) = \boxed{1|0|\cdots|0|0|1|\cdots|0|\cdots|0|0|\cdots|1} \in \{0,1\}^{Lq}$
q
q
q

**Fig 2. Illustration of binary embedding.** The binary embedding $\mathbf{1}_\sigma$: $\Omega \rightarrow \{0, 1\}^{Lq}$ maps each vector of categorical random variables, $\mathbf{x} \in \Omega^L$, here represented by a sequence of amino acids from the amino acid alphabet (containing the 20 amino acids and one gap element), $\Omega = \{A, C, D, E, F, G, H, I, K, L, M, N, P, Q, R, S, T, V, W, Y, -\}$, onto a unique binary representation, $\mathbf{x}(\boldsymbol{\sigma}) \in \{0, 1\}^{Lq}$.

doi:10.1371/journal.pcbi.1004182.g002

we use a so-called multiple sequence alignment, $\{\mathbf{x}^1,\ldots, \mathbf{x}^M\} \subset \Omega^{L\times M}$, a collection of closely homologous protein sequences that is formatted such that it allows comparison of the evolution across each residue [44]. These alignments may stem from different hidden Markov model-derived resources, such as PFAM [45], hhblits [46], and Jackhmmer [47].

To formalize the derivation of the pairwise maximum-entropy probability distribution on categorical variables, we use the approach of [8,30,48] and replace, as depicted in Fig 2, each variable $x_i$ defined on categorical variables by an indicator function of the amino acid $\sigma \in \Omega$, $\mathbf{1}_\sigma$: $\Omega \rightarrow \{0, 1\}^q$,

$$x_i \mapsto x_i(\sigma) :\equiv \mathbf{1}_\sigma(x_i) = \begin{cases} 1 & \text{if } x_i = \sigma, \\ 0 & \text{otherwise.} \end{cases}$$

This embedding specifies a unique representation of any $L$-vector of categorical random variables, $\mathbf{x}$, as a binary $Lq$-vector, $\mathbf{x}(\boldsymbol{\sigma})$ with a single non-zero entry in each binary $q$-subvector $x_i(\boldsymbol{\sigma}) = (x_i(\sigma_1),\ldots, x_i(\sigma_q))^{\mathrm{T}} \in \{0,1\}^q$,

$$\mathbf{x} = (x_1,\ldots,x_L)^{\mathrm{T}} \in \Omega^L \overset{\mathbf{1}_\sigma}{\mapsto} \mathbf{x}(\boldsymbol{\sigma}) = (x_1(\sigma_1),\ldots,x_L(\sigma_q))^{\mathrm{T}} \in \{0,1\}^{Lq}.$$

Inserting this embedding into the first and second moment constraints, corresponding to Eqs 3 and 4 in the continuous variable case, we find their embedded analogues, the single and pairwise marginal probability in positions $i$ and $j$ for amino acids $\sigma,\omega,\in\Omega$

$$\langle x_i(\sigma)\rangle = \sum_{\mathbf{x}(\boldsymbol{\sigma})} P(\mathbf{x}(\boldsymbol{\sigma}))x_i(\sigma) = \sum_{\mathbf{x}} P(x_i = \sigma) = P_i(\sigma),$$

$$\langle x_i(\sigma)x_j(\omega)\rangle = \sum_{\mathbf{x}(\boldsymbol{\sigma})} P(\mathbf{x}(\boldsymbol{\sigma}))x_i(\sigma)x_j(\omega) = \sum_{\mathbf{x}} P(x_i = \sigma, x_j = \omega) = P_{ij}(\sigma,\omega)$$

including $P_{ii}(\sigma,\omega) = P_i(\sigma)\mathbf{1}_\sigma(\omega)$ and with the distribution's first moment in each random variable, $\langle y_i\rangle = \sum_{\mathbf{y}} P(\mathbf{y})y_i$ and $\boldsymbol{y} = (y_1,\ldots, y_{Lq})^{\mathrm{T}} \in \mathbb{R}^{Lq}$. The analogue of the covariance matrix then becomes a symmetric $Lq \times Lq$ matrix of connected correlations whose entries $C_{ij}(\sigma,\omega) = P_{ij}(\sigma,\omega) - P_i(\sigma)\, P_j(\omega)$ characterize the dependencies between pairs of variables. In the same way, the sample means translate to the single-site and pair frequency counts over $m = 1,\ldots,M$ data vectors $\mathbf{x}^m = (x_1^m, \ldots, x_L^m)^{\mathrm{T}} \in \mathbf{\Omega}^L$,

$$\overline{x_i(\sigma)} = \frac{1}{M}\sum_{m=1}^{M} x_i^m(\sigma) = f_i(\sigma),$$

$$\overline{x_i(\sigma)x_j(\omega)} = \frac{1}{M}\sum_{m=1}^{M} x_i^m(\sigma)x_j^m(\omega) = f_{ij}(\sigma,\omega).$$

The pairwise maximum-entropy probability distribution in categorical variables has to fulfill the normalization constraint,

$$\sum_{\mathbf{x}} P(\mathbf{x}) = \sum_{\mathbf{x}(\boldsymbol{\sigma})} P(\mathbf{x}(\boldsymbol{\sigma})) = 1. \tag{9}$$

Furthermore, the single and pair constraints, the analogues of Eqs 3 and 4, enforce the resulting probability distribution to be compatible with the measured single and pair frequency counts,

$$P_i(\sigma) = f_i(\sigma), \qquad P_{ij}(\sigma,\omega) = f_{ij}(\sigma,\omega) \tag{10}$$

for each $i, j = 1,\ldots, L$ and amino acids $\sigma,\omega \in \Omega$. As before, we require the probability distribution to maximize the information entropy,

$$\text{maximize } S = -\sum_{\mathbf{x}} P(\mathbf{x}) \ln P(\mathbf{x}) = -\sum_{\mathbf{x}(\boldsymbol{\sigma})} P(\mathbf{x}(\boldsymbol{\sigma})) \ln P(\mathbf{x}(\boldsymbol{\sigma})). \tag{11}$$

The corresponding Lagrangian, $\mathcal{L} = \mathcal{L}(P(\mathbf{x}(\boldsymbol{\sigma})); \alpha, \boldsymbol{\beta}(\boldsymbol{\sigma}), \boldsymbol{\gamma}(\boldsymbol{\sigma},\boldsymbol{\omega}))$, has the functional form,

$$\mathcal{L} = S + \alpha(\langle 1 \rangle - 1) + \sum_{i=1}^{L}\sum_{\sigma\in\Omega} \beta_i(\sigma)(P_i(\sigma) - f_i(\sigma)) + \sum_{i,j=1}^{L}\sum_{\sigma,\omega\in\Omega} \gamma_{ij}(\sigma,\omega)(P_{ij}(\sigma,\omega) - f_{ij}(\sigma,\omega)).$$

For notational convenience, the Lagrange multipliers $\beta_i(\sigma)$ and $\gamma_{ij}(\sigma,\omega)$ are grouped to the $Lq$-vector $\boldsymbol{\beta}(\boldsymbol{\sigma}) = (\beta_i(\sigma))_{i=1,\ldots,L}^{\sigma\in\Omega}$ and the $Lq \times Lq$-matrix $\boldsymbol{\gamma}(\boldsymbol{\sigma},\boldsymbol{\omega}) = (\gamma_{ij}(\sigma,\omega))_{i,j=1,\ldots,L}^{\sigma,\omega\in\Omega}$, respectively. The Lagrangian's stationary point, found as the solution of $\frac{\partial\mathcal{L}}{\partial P(\mathbf{x}(\boldsymbol{\sigma}))} = 0$, determines the pairwise maximum-entropy probability distribution in categorical variables [30,49],

$$P(\mathbf{x}(\boldsymbol{\sigma}); \boldsymbol{\beta}, \boldsymbol{\gamma}) = \frac{1}{Z}\exp\left(\sum_{i=1}^{L}\sum_{\sigma\in\Omega} \beta_i(\sigma)x_i(\sigma) + \sum_{i,j=1}^{L}\sum_{\sigma,\omega\in\Omega} \gamma_{ij}(\sigma,\omega)x_i(\sigma)x_j(\omega)\right) \tag{12}$$

with normalization by the partition function, $Z \equiv \exp(1-\alpha)$. Note that distribution Eq 12 is of the same functional form as Eq 8 but with binary random variables $\mathbf{x}(\boldsymbol{\sigma}) \in \{0,1\}^{Lq}$ instead of continuous ones $\mathbf{x} \in \mathbb{R}^L$. At this point, we introduce the reduced parameter set, $h_i(\sigma) := \beta_i(\sigma) + \gamma_{ii}(\sigma, \sigma)$ and $e_{ij}(\sigma,\omega) := 2\gamma_{ij}(\sigma,\omega)$ for $i < j$, using the symmetry of the Lagrange multipliers, $\gamma_{ij}(\sigma,\omega) := \gamma_{ji}(\omega, \sigma)$, and that $x_i(\sigma)\, x_i(\omega) = 1$ if and only if $\sigma = \omega$. For a given sequence $(z_1,\ldots, z_L) \in \Omega^L$ summing over all non-zero elements, $(x_1(z_1) = 1,\ldots, x_L(z_L) = 1)$ or equivalently $(x_1 = z_1,\ldots,x_L = z_L)$ then yields the probability assigned to the sequence of interest,

$$P(z_1,\ldots,z_L) \equiv \frac{1}{Z}\exp\left(\sum_{i=1}^{L} h_i(z_i) + \sum_{1\le i<j\le L} e_{ij}(z_i,z_j)\right). \tag{13}$$

This is the 21-state maximum-entropy probability distribution as presented by [5–7].

## Gauge fixing

In contrast to the continuous variable case in which the number of constraints naturally matches the number of unknown parameters, the case of categorical variables has dependencies due to $1 = \sum_{\sigma\in\Omega} P_i(\sigma)$ for each $i = 1,\ldots, L$ and $P_i(\sigma) = \sum_{\omega\in\Omega} P_{ij}(\sigma,\omega)$ for each $i, j = 1,\ldots, L$ and $\sigma\in\Omega$. This results in at most $\frac{L(L-1)}{2}(q-1)^2 + L(q-1)$ independent constraints compared to $\frac{L(L-1)}{2}q^2 + Lq$ free parameters to be estimated. To ensure the uniqueness of the inferred parameters in defining the Hamiltonian, $\mathcal{H}(x_1,\ldots,x_L) = -\sum_{i<j} e_{ij}(x_i,x_j) - \sum_i h_i(x_i)$, and, by implication, the probability distribution, one has to reduce the number of independent parameters such that these match the number of independent constraints. For this purpose, so-called gauge fixing [5] has been proposed, which can be realized in different ways. For example, the authors of [6,7] set the parameters corresponding to the last amino acid in the alphabet, $\sigma_q$, to zero, i.e., $e_{ij}(\sigma_q,\cdot) = e_{ij}(\cdot,\sigma_q) = 0$ and $h_i(\sigma_q) = 0$ for $1 \leq i < j \leq L$, resulting in rows and columns of zeros at the end of each $q \times q$-block of the $Lq \times Lq$ coupling matrix. Alternatively, the authors of [5] introduce a zero-sum gauge, $\sum_\sigma e_{ij}(\sigma,\omega) = \sum_\sigma e_{ij}(\omega',\sigma) = 0$ and $\sum_\sigma h_i(\sigma) = 0$ for each $1 \leq i < j \leq L$ and $\omega, \omega\in\Omega$. However, different gauge fixings are not equally efficient for the purpose of protein contact prediction. The zero-sum gauge is the parameter fixing that minimizes the sum of squares of the pairwise parameters in the Hamiltonian $\mathcal{H}$, $\sum_{\sigma,\omega} e_{ij}(\sigma,\omega)^2$, which makes it the suitable choice when using non-gauge invariant scoring functions, such as the (average product-corrected) Frobenius norm [5,50] (see section "Scoring Functions"). Moreover, no gauge fixing is required when combining the strictly convex $\ell^1$- or $\ell^2$-regularizer with negative loglikelihood minimization; here the regularizer selects for a unique representation among all parametrizations of the optimal distribution [32,51]. However, to additionally minimize the Frobenius norm of the pairwise interactions, [51] changed the obtained full parameter set from regularized inference with plmDCA to zero-sum gauge by, $e_{ij}(\sigma,\omega) \mapsto e_{ij}(\sigma,\omega) - \frac{1}{q}\sum_{\sigma'} e_{ij}(\sigma',\omega) - \frac{1}{q}\sum_{\omega'} e_{ij}(\sigma,\omega') + \frac{1}{q^2}\sum_{\sigma',\omega'} e_{ij}(\sigma',\omega')$, where $q$ denotes the length of the alphabet.

## Network interpretation

The derived pairwise maximum-entropy distributions in Eqs 13 or 12 and 8 specify an undirected graphical model or Markov random field [34,41]. In particular, a graphical model represents a probability distribution in terms of a graph that consists of a node and an edge set. Edges characterize the dependence structure between nodes and a missing edge then corresponds to **conditional independence** given the remaining random variables. For continuous, real-valued variables, the maximum-entropy distribution with first and second moment constraints is multivariate Gaussian, which will be demonstrated in the next section. Its dependency structure is represented by a graphical Gaussian model (GGM) in which a missing edge, $\gamma_{ij} = 0$, corresponds to conditional independence between the random variables $x_i$ and $x_j$ (given the remaining ones), and is further specified by a zero entry in the corresponding inverse covariance matrix, $(C^{-1})_{ij} = 0$.

In the next section, we describe how the dependency structure of the graph is inferred.

# Inference of Interactions

Up to this point, the functional form of the maximum-entropy probability distribution is specified, but not its determining parameters. For categorical variables with dimension $L > 1$, there is typically no closed-form solution. In the following section, we present several inference methods to estimate these parameters that have recently been used in the context of protein contact prediction. Those are (1) for continuous variables, the exact closed-form solution which approximates the mean-field result for categorical variables, and (2) three inference methods for categorical variables based on the maximum-likelihood methodology: the stochastic maximum likelihood, the approximation by pseudo-likelihood maximization, and finally, the sparse maximum-likelihood solution.

## Closed-Form Solution for Continuous Variables

The simplest approach to extract the unknown Lagrange multipliers $\alpha$, $\boldsymbol{\beta} = (\beta_i)$, and $\boldsymbol{\gamma} = (\gamma_{ij})$ from $P(\mathbf{x})$ exactly is to use basic integration properties of the continuous random variables $x_i$ in the constraints Eqs 3–5. For this purpose, we rewrite the exponent of the pairwise maximum-entropy probability distribution Eq 8,

$$P(\mathbf{x};\boldsymbol{\beta},\tilde{\boldsymbol{\gamma}}) = \frac{1}{Z}\exp\left(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x} - \frac{1}{2}\mathbf{x}^{\mathrm{T}}\tilde{\boldsymbol{\gamma}}\mathbf{x}\right) = \frac{1}{Z}\exp\left(\frac{1}{2}\boldsymbol{\beta}^{\mathrm{T}}\tilde{\boldsymbol{\gamma}}^{-1}\boldsymbol{\beta} - \frac{1}{2}(\mathbf{x}-\tilde{\boldsymbol{\gamma}}^{-1}\boldsymbol{\beta})^{\mathrm{T}}\tilde{\boldsymbol{\gamma}}(\mathbf{x}-\tilde{\boldsymbol{\gamma}}^{-1}\boldsymbol{\beta})\right),$$

where we use the replacement $\tilde{\boldsymbol{\gamma}} := -2\boldsymbol{\gamma}$ and require $\tilde{\boldsymbol{\gamma}}$ to be positive definite (which is equivalent to $\boldsymbol{\gamma}$ being negative definite), i.e., $\mathbf{x}^{\mathrm{T}}\tilde{\boldsymbol{\gamma}}\mathbf{x} > 0$ for any $\mathbf{x} \neq \mathbf{0}$, which makes its inverse $\tilde{\boldsymbol{\gamma}}^{-1} = -\frac{1}{2}\boldsymbol{\gamma}^{-1}$ well-defined. As already discussed, this is a sufficient condition on the integrals in Eqs 3–6 to be finite. For notational convenience, we define the shifted variable $\mathbf{z} = (z_1, \ldots, z_L)^{\mathrm{T}} := \mathbf{x} - \tilde{\boldsymbol{\gamma}}^{-1}\boldsymbol{\beta}$ or $x_i = z_i + \sum_{j=1}^{L}(\tilde{\boldsymbol{\gamma}}^{-1})_{ij}\beta_j$ and accordingly, the maximum-entropy probability distribution becomes

$$P(\mathbf{x}) = \frac{1}{\tilde{Z}}\exp\left(-\frac{1}{2}(\mathbf{x}-\tilde{\boldsymbol{\gamma}}^{-1}\boldsymbol{\beta})^{\mathrm{T}}\tilde{\boldsymbol{\gamma}}(\mathbf{x}-\tilde{\boldsymbol{\gamma}}^{-1}\boldsymbol{\beta})\right) \equiv \frac{1}{\tilde{Z}}\mathrm{e}^{-\frac{1}{2}\mathbf{z}^{\mathrm{T}}\tilde{\boldsymbol{\gamma}}\mathbf{z}} \qquad (14)$$

with the normalization constant $\tilde{Z} = \exp\left(1 - \alpha - \frac{1}{2}\boldsymbol{\beta}^{\mathrm{T}}\tilde{\boldsymbol{\gamma}}^{-1}\boldsymbol{\beta}\right)$. The normalization condition Eq 3 in the new variable is,

$$1 = \int_{\mathbf{x}} P(\mathbf{x})\,\mathrm{d}\mathbf{x} \equiv \frac{1}{\tilde{Z}}\int_{\mathbf{z}} \mathrm{e}^{-\frac{1}{2}\mathbf{z}^{\mathrm{T}}\tilde{\boldsymbol{\gamma}}\mathbf{z}}\,\mathrm{d}\mathbf{z} \qquad (15)$$

and the linear shift does not affect the integral when integrated over $\mathbb{R}^L$ yielding for the normalization constant, $\tilde{Z} = \int_{\mathbf{z}} \mathrm{e}^{-\frac{1}{2}\mathbf{z}^{\mathrm{T}}\tilde{\boldsymbol{\gamma}}\mathbf{z}}\,\mathrm{d}\mathbf{z}$. Furthermore, the first-order constraint Eq 4 becomes for each $i = 1,\ldots, L$,

$$\langle x_i\rangle = \int_{\mathbf{x}} P(\mathbf{x})x_i\,\mathrm{d}\mathbf{x} \equiv \frac{1}{\tilde{Z}}\int_{\mathbf{z}} \mathrm{e}^{-\frac{1}{2}\mathbf{z}^{\mathrm{T}}\tilde{\boldsymbol{\gamma}}\mathbf{z}}\left(z_i + \sum_{j=1}^{L}(\tilde{\boldsymbol{\gamma}}^{-1})_{ij}\beta_j\right)\mathrm{d}\mathbf{z} = \sum_{j=1}^{L}(\tilde{\boldsymbol{\gamma}}^{-1})_{ij}\beta_j$$

and we used the point symmetry of the integrand then, $\int_{\mathbf{z}} \mathrm{e}^{-\frac{1}{2}\mathbf{z}^{\mathrm{T}}\tilde{\boldsymbol{\gamma}}\mathbf{z}} z_i\,\mathrm{d}\mathbf{z} = 0$ in each $i = 1,\ldots, L$. Analogously, we find for the second moment, determining the correlations for each index pair $i, j = 1,\ldots, L$,

$$\langle x_i x_j\rangle = \int_{\mathbf{x}} P(\mathbf{x})x_i x_j\,\mathrm{d}\mathbf{x} \equiv \frac{1}{\tilde{Z}}\int_{\mathbf{z}} \mathrm{e}^{-\frac{1}{2}\mathbf{z}^{\mathrm{T}}\tilde{\boldsymbol{\gamma}}\mathbf{z}}(z_i - \langle x_i\rangle)(z_j - \langle x_j\rangle)\,\mathrm{d}\mathbf{z} = \langle z_i z_j\rangle + \langle x_i\rangle\langle x_j\rangle,$$

where we use again the point symmetry and the result on the normalization constraint. Based on this, the covariance is found as,

$$C_{ij} = \langle x_i x_j \rangle - \langle x_i \rangle \langle x_j \rangle \equiv \langle z_i z_j \rangle.$$

Finally, the term $\langle z_i\, z_j \rangle$ is solved using a spectral decomposition of the symmetric and positive-definite matrix $\tilde{\boldsymbol{\gamma}}$ as sum over products of its eigenvectors $\mathbf{v}_1,\ldots,\mathbf{v}_L$ and real-valued and positive eigenvalues $\lambda_1,\ldots,\lambda_L$, $\tilde{\boldsymbol{\gamma}} = \sum_{k=1}^{L} \lambda_k \mathbf{v}_k \mathbf{v}_k^{\mathrm{T}}$. The eigenvectors form a basis of $\mathbb{R}^L$ and assign new coordinates, $y_1,\ldots,y_L$, to $\mathbf{z} = \sum_{k=1}^{L} y_k \mathbf{v}_k$, which allows writing of the exponent $\langle z_i\, z_j \rangle$ as $\mathbf{z}^{\mathrm{T}} \tilde{\boldsymbol{\gamma}} \mathbf{z} = \sum_{k=1}^{L} \lambda_k y_k^2$. The covariance between $x_i$ and $x_j$ then reads as (Bishop [52], p. 83)

$$\langle z_i z_j \rangle = \frac{1}{\tilde{Z}} \sum_{l,n=1}^{L} (v_l)_i (v_n)_j \int_{\mathbf{y}} \exp\left( -\frac{1}{2} \sum_{k=1}^{L} \lambda_k y_k^2 \right) y_l y_n \, \mathrm{d}\mathbf{y} = \sum_{k=1}^{L} \frac{1}{\lambda_k} (\mathbf{v}_k)_i (\mathbf{v}_k)_j \equiv (\tilde{\boldsymbol{\gamma}}^{-1})_{ij}$$

with solution $C_{ij} = (\tilde{\boldsymbol{\gamma}}^{-1})_{ij}$ or $(C^{-1})_{ij} = (\tilde{\boldsymbol{\gamma}})_{ij} = -2\gamma_{ij}$. Taken together, the Lagrange multipliers $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ are specified in terms of the mean, $\langle \mathbf{x} \rangle$, and the inverse covariance matrix (also known as the precision or concentration matrix), $C^{-1}$,

$$\boldsymbol{\beta} = C^{-1} \langle \mathbf{x} \rangle, \qquad \boldsymbol{\gamma} = -\frac{1}{2} \tilde{\boldsymbol{\gamma}} = -\frac{1}{2} C^{-1}. \tag{16}$$

As a consequence, the real-valued maximum-entropy distribution Eq 14 for given first and second moments is found as the **multivariate Gaussian distribution,** which is determined by the mean $\langle \mathbf{x} \rangle$ and the covariance matrix $C$,

$$P(\mathbf{x}; \langle \mathbf{x} \rangle, C) = (2\pi)^{-L/2} \det(C)^{-1/2} \exp\left( -\frac{1}{2} (\mathbf{x} - \langle \mathbf{x} \rangle)^{\mathrm{T}} C^{-1} (\mathbf{x} - \langle \mathbf{x} \rangle) \right) \tag{17}$$

and we refer to [52] for the derivation of the normalization factor. The initial requirement of $\tilde{\boldsymbol{\gamma}} = -2\boldsymbol{\gamma}$ to be positive definite results in a positive-definite covariance matrix $C$, a necessary condition for the Gaussian density to be well defined. In summary, the multivariate Gaussian distribution maximizes the entropy among all probability distributions of continuous variables with specified first and second moments. The pair interaction strength is now evaluated by the already introduced partial correlation coefficient between $x_i$ and $x_j$ given the remaining variables $\{x_r\}_{r \in \{1,\ldots,L\} \setminus \{i,j\}}$,

$$\rho_{ij\cdot\{1,\ldots,L\}\setminus\{i,j\}} \equiv \frac{\gamma_{ij}}{\sqrt{\gamma_{ii}\gamma_{jj}}} = \begin{cases} -\dfrac{(C^{-1})_{ij}}{\sqrt{(C^{-1})_{ii}(C^{-1})_{jj}}} & \text{if } i \neq j, \\ 1 & \text{if } i = j. \end{cases} \tag{18}$$

## Data integration

In biological datasets as used to study gene association, the number of measurements, $M$, is typically smaller than the number of observables, $L$, i.e., $M < L$ in our terminology. Consequently, the empirical covariance matrix, $\hat{C} = \frac{1}{M} \sum_{m=1}^{M} (\mathbf{x}^m - \overline{\mathbf{x}})(\mathbf{x}^m - \overline{\mathbf{x}})^{\mathrm{T}}$, will in these cases always be rank-deficient (and, thus, not invertible) since its rank can exceed neither the number of variables, $L$, nor the number of measurements, $M$. Moreover, even in cases when $M \geq L$, the empirical covariance matrix may become non-invertible or badly conditioned (i.e., close to singular) due to dependencies in the data. However, for variables following a multivariate Gaussian distribution, one can access the elements of its inverse by maximizing the penalized Gaussian loglikelihood, which results in the following estimate of the inverse covariance matrix, $C^{-1} \approx C^{-1}_{\delta,\lambda}$,

$$C^{-1}_{\delta,\lambda} = \underset{\substack{\Theta \text{ pos. definite,} \\ \text{symmetric}}}{\arg\max}\{\ln\det(\Theta) - \text{trace}(\hat{C}\Theta) - \lambda\|\Theta\|^{\delta}_{\delta}\} \tag{19}$$

with penalty parameter $\lambda \geq 0$ and $\|\Theta\|^{\delta}_{\delta} = \sum_{i,j}|\Theta_{ij}|^{\delta}$. If $\lambda = 0$, we obtain the maximum-likelihood estimate, for $\delta = 1$ and $\lambda > 0$ the $\ell^1$-regularized (sparse) maximum-likelihood solution that selects for sparsity [53,54], and for $\delta = 2$ and $\lambda > 0$ the $\ell^2$-regularized maximum-likelihood solution that favors small absolute values in the entries of the selected inverse covariance matrix [55]. For $\delta = 1$ and $\lambda > 0$, the method is called LASSO, for $\delta = 2$ and $\lambda > 0$, ridge regression. Alternatively, regularization can be directly applied to the covariance matrix, e.g., by shrinkage [17,56].

## Solution for categorical variables

An ad hoc ansatz to extract the pairwise parameters in the categorical variables case (12) is to extend the binary variable $\mathbf{x}(\boldsymbol{\sigma}) = (x_i(\sigma_k))_{i\cdot k} \in \{0,1\}^{L(q-1)}$ to a continuous one, $\mathbf{y} = (y_j)_j \in \mathbb{R}^{L(q-1)}$, and replace the sums in the distribution and the moments $\langle\cdot\rangle$ by integrals. The extended binary maximum-entropy distribution Eq 12 is then approximated by the $Lq$-dimensional multivariate Gaussian with inherited analogues of the mean $\langle\mathbf{y}\rangle = (f_i(\sigma_k))_{i\cdot k} \in \mathbb{R}^{L(q-1)}$ and the empirical covariance matrix $\hat{C}(\boldsymbol{\sigma},\boldsymbol{\omega}) = (\hat{C}_{ij}(\sigma_k,\sigma_l))_{i,j,k,l} \in \mathbb{R}^{L(q-1)\times L(q-1)}$ whose elements $\hat{C}_{ij}(\sigma,\omega) = f_{ij}(\sigma,\omega) - f_i(\sigma)f_j(\omega)$ are characterizing the pairwise dependency structure. The gauge fixing results in setting the preassigned entries referring to the last amino acid in the mean vector and the covariance matrix to zero, which reduces the model's dimension from $Lq$ to $L(q-1)$; otherwise the unregularized covariance matrix would always be non-invertible. Typically, the single and pair frequency counts are reweighted and regularized by pseudocounts (see section "Sequence data preprocessing") to additionally ensure that $\hat{C}(\boldsymbol{\sigma},\boldsymbol{\omega})$ is invertible. Final application of the closed-form solution for continuous variables Eq 16 to the extended binary variables for $C^{-1}(\boldsymbol{\sigma},\boldsymbol{\omega}) \approx \hat{C}^{-1}(\boldsymbol{\sigma},\boldsymbol{\omega})$ yields the so-called mean-field (MF) approximation [48],

$$\gamma^{\text{MF}}_{ij}(\sigma,\omega) = -\frac{1}{2}(C^{-1})_{ij}(\sigma,\omega) \quad \Rightarrow \quad e^{\text{MF}}_{ij}(\sigma,\omega) = -(C^{-1})_{ij}(\sigma,\omega) \tag{20}$$

for amino acids $\sigma,\omega \in \Omega$ and with restriction to residues $i < j$ in the latter identity. The same solution has been obtained by [6,7] using a perturbation ansatz to solve the $q$-state Potts model termed (mean-field) Direct Coupling Analysis (DCA or mfDCA). In Ising models, this result is also known as naïve mean-field approximation [57–59].

The following section is dedicated to maximum likelihood-based inference approaches, which have been presented in the field of protein contact prediction.

# Maximum-Likelihood Inference

A well-known approach to estimate the parameters of a model is maximum-likelihood inference. The likelihood is a scalar measure of how likely the model parameters are, given the observed data (Mackay [34], p. 29), and the maximum-likelihood solution denotes the parameter set maximizing the likelihood function. For Markov random fields, the maximum-likelihood solution is consistent, i.e., recovers the true model parameters in the limit of infinite data (Koller and Friedman [32], p. 949). In particular, for a pairwise model with parameters $\boldsymbol{h}(\boldsymbol{\sigma}) = (h_i(\sigma))^{\sigma\in\Omega}_{i=1,\ldots,L}$ and $\boldsymbol{e}(\boldsymbol{\sigma},\boldsymbol{\omega}) = (e_{ij}(\sigma,\omega))^{\sigma,\omega\in\Omega}_{1\leq i<j\leq L}$, we find the likelihood $l(\boldsymbol{h}(\boldsymbol{\sigma}),\boldsymbol{e}(\boldsymbol{\sigma},\boldsymbol{\omega})) = l(\boldsymbol{h}(\boldsymbol{\sigma}),\boldsymbol{e}(\boldsymbol{\sigma},$

$\boldsymbol{\omega})|\mathbf{x}^1,\ldots,\mathbf{x}^M)$ given observed data, $\mathbf{x}^1,\ldots,\mathbf{x}^M \in\Omega^L$,which are assumed to be independent and identically distributed (iid), as

$$l(\boldsymbol{h}(\boldsymbol{\sigma}),\boldsymbol{e}(\boldsymbol{\sigma},\boldsymbol{\omega})|\mathrm{x}^1,\ldots,\mathrm{x}^M) = \prod_{m=1}^{M} P(\mathrm{x}^m;\boldsymbol{h}(\boldsymbol{\sigma}),\boldsymbol{e}(\boldsymbol{\sigma},\boldsymbol{\omega})). \tag{21}$$

The estimates of the model parameters are then obtained as the maximizer of l or, using the monotonicity of the logarithm, the minimizer of ln l,

$$\{\boldsymbol{h}^{\mathrm{ML}}(\boldsymbol{\sigma}),\boldsymbol{e}^{\mathrm{ML}}(\boldsymbol{\sigma},\boldsymbol{\omega})\} = \underset{h(\sigma),e(\sigma,\omega)}{\arg\max}\ l(\boldsymbol{h}(\boldsymbol{\sigma}),\boldsymbol{e}(\boldsymbol{\sigma},\boldsymbol{\omega})) \equiv \underset{h(\sigma),e(\sigma,\omega)}{\arg\min}\ -\ln l(\boldsymbol{h}(\boldsymbol{\sigma}),\boldsymbol{e}(\boldsymbol{\sigma},\boldsymbol{\omega})).$$

When we specify the maximum-entropy distribution Eq 13 as model distribution, the then-concave loglikelihood [32] becomes

$$\begin{aligned}\ln l(\boldsymbol{h}(\boldsymbol{\sigma}),\boldsymbol{e}(\boldsymbol{\sigma},\boldsymbol{\omega})) &= \sum_{m=1}^{M}\ln P(\mathrm{x}^m;\boldsymbol{h}(\boldsymbol{\sigma}),\boldsymbol{e}(\boldsymbol{\sigma},\boldsymbol{\omega}))\\ &= -M\left[\ln Z - \sum_{i=1}^{L}\sum_{\sigma} h_i(\sigma)f_i(\sigma) - \sum_{1\le i<j\le L}\sum_{\sigma,\omega} e_{ij}(\sigma,\omega)f_{ij}(\sigma,\omega)\right].\end{aligned} \tag{22}$$

The maximum-likelihood solution is found by taking the derivatives of Eq 22 with respect to the model parameters $h_i(\sigma)$ and $e_{ij}(\sigma,\omega)$ and setting to zero,

$$\begin{aligned}\frac{\partial}{\partial h_i(\sigma)}\ln l &= -M\left[\frac{\partial}{\partial h_i(\sigma)}\ln Z\Big|_{\{\boldsymbol{h}(\boldsymbol{\sigma}),\boldsymbol{e}(\boldsymbol{\sigma},\boldsymbol{\omega})\}} - f_i(\sigma)\right] = 0,\\ \frac{\partial}{\partial e_{ij}(\sigma,\omega)}\ln l &= -M\left[\frac{\partial}{\partial e_{ij}(\sigma,\omega)}\ln Z\Big|_{\{\boldsymbol{h}(\boldsymbol{\sigma}),\boldsymbol{e}(\boldsymbol{\sigma},\boldsymbol{\omega})\}} - f_{ij}(\sigma,\omega)\right] = 0.\end{aligned} \tag{23}$$

The partial derivatives of the partition function, $Z = \sum_{(x_1,\ldots,x_L)}\exp\left(\sum_i h_i(x_i) + \sum_{i<j} e_{ij}(x_i,x_j)\right)$, follow the well-known identities

$$\frac{\partial}{\partial h_i(\sigma)}\ln Z\Big|_{\{\boldsymbol{h}(\boldsymbol{\sigma}),\boldsymbol{e}(\boldsymbol{\sigma},\boldsymbol{\omega})\}} = \frac{1}{Z}\partial_{h_i(\sigma)}Z\Big|_{\{\boldsymbol{h}(\boldsymbol{\sigma}),\boldsymbol{e}(\boldsymbol{\sigma},\boldsymbol{\omega})\}} = P_i(\sigma;\boldsymbol{h}(\boldsymbol{\sigma}),\boldsymbol{e}(\boldsymbol{\sigma},\boldsymbol{\omega})),$$

$$\frac{\partial}{\partial e_{ij}(\sigma,\omega)}\ln Z\Big|_{\{\boldsymbol{h}(\boldsymbol{\sigma}),\boldsymbol{e}(\boldsymbol{\sigma},\boldsymbol{\omega})\}} = \frac{1}{Z}\partial_{e_{ij}(\sigma,\omega)}Z\Big|_{\{\boldsymbol{h}(\boldsymbol{\sigma}),\boldsymbol{e}(\boldsymbol{\sigma},\boldsymbol{\omega})\}} = P_{ij}(\sigma,\omega;\boldsymbol{h}(\boldsymbol{\sigma}),\boldsymbol{e}(\boldsymbol{\sigma},\boldsymbol{\omega})).$$

The maximizing parameters, $\boldsymbol{h}^{\mathrm{ML}}(\boldsymbol{\sigma}) = (h_i^{\mathrm{ML}}(\sigma))_{i=1,\ldots,L}^{\sigma\in\Omega}$ and $\boldsymbol{e}^{\mathrm{ML}}(\boldsymbol{\sigma},\boldsymbol{\omega}) = (e_{ij}^{\mathrm{ML}}(\sigma,\omega))_{1\le i<j\le L}^{\sigma,\omega\in\Omega}$, are those matching the distribution's single and pair marginal probabilities with the empirical single and pair frequency counts,

$$P_i(\sigma;\boldsymbol{h}^{\mathrm{ML}}(\boldsymbol{\sigma}),\boldsymbol{e}^{\mathrm{ML}}(\boldsymbol{\sigma},\boldsymbol{\omega})) = f_i(\sigma), \qquad P_{ij}(\sigma,\omega;\boldsymbol{h}^{\mathrm{ML}}(\boldsymbol{\sigma}),\boldsymbol{e}^{\mathrm{ML}}(\boldsymbol{\sigma},\boldsymbol{\omega})) = f_{ij}(\sigma,\omega)$$

in residues $i = 1,\ldots, L$ and $i,j = 1,\ldots, L$, respectively, and for amino acids $\sigma,\omega\in\Omega$. In other words, matching the moments of the pairwise maximum-entropy probability distribution to the given data is equivalent to maximum-likelihood fitting of an exponential family [34,60]. Although the maximum-likelihood solution is globally optimal for the pairwise maximum-entropy probability model, based on the concavity of ln $l$, the resulting distribution is not necessarily unique, due to dependencies in the input data (Koller and Friedman [32], p. 948). To remove these equivalent optima and select for a unique representation, one needs to introduce further constraints by, for example, gauge fixing or regularization.

Based on the maximum-likelihood principle, we present three solution approaches in the remainder of this section.

## Stochastic maximum likelihood

The maximum-likelihood solution is typically inaccessible for models of categorical variables due to the computational complexity of estimating the partition function $Z$ which involves a sum over all possible states and grows exponentially with the size of the system [3,61]. Lapedes et al. [30] solved Eq 22 by likelihood maximization on sampled subsets using the Metropolis–Hastings algorithm [32,34]. In particular, the likelihood is maximized iteratively by following the steepest ascent of the loglikelihood function ln $l$ using Eq 23. In each maximization step, the parameters $h_i^{(k)}(\sigma)$ and $e_{ij}^{(k)}(\sigma,\omega)$ are changed in proportion to the gradient of ln $l$ and scaled by the constant step size $\varepsilon > 0$,

$$\Delta h_i^{(k)}(\sigma) = \varepsilon \frac{\partial}{\partial h_i(\sigma)} \ln l \Big|_{\{\boldsymbol{h}^{(k)}(\boldsymbol{\sigma}),\boldsymbol{e}^{(k)}(\boldsymbol{\sigma},\boldsymbol{\omega})\}} \propto f_i(\sigma) - P_i(\sigma;\boldsymbol{h}^{(k)}(\boldsymbol{\sigma}),\boldsymbol{e}^{(k)}(\boldsymbol{\sigma},\boldsymbol{\omega})),$$

$$\Delta e_{ij}^{(k)}(\sigma,\omega) = \varepsilon \frac{\partial}{\partial e_{ij}(\sigma,\omega)} \ln l \Big|_{\{\boldsymbol{h}^{(k)}(\boldsymbol{\sigma}),\boldsymbol{e}^{(k)}(\boldsymbol{\sigma},\boldsymbol{\omega})\}} \propto f_{ij}(\sigma,\omega) - P_{ij}(\sigma,\omega;\boldsymbol{h}^{(k)}(\boldsymbol{\sigma}),\boldsymbol{e}^{(k)}(\boldsymbol{\sigma},\boldsymbol{\omega}))$$

until convergence is reached as the differences $\Delta h_i^{(k)}(\sigma,\omega) := h_i^{(k+1)}(\sigma,\omega) - h_i^{(k)}(\sigma,\omega)$, $i = 1,\ldots, L$, and $\Delta e_{ij}^{(k)}(\sigma,\omega) := e_{ij}^{(k+1)}(\sigma,\omega) - e_{ij}^{(k)}(\sigma,\omega)$, $1 \le i < j \le L$, go to zero [30]. The computation of the marginals requires summing over $20^L$ states and is, for example, estimated by Monte-Carlo sampling. As the likelihood is concave, there are no local maxima and the maximum-likelihood parameters are obtained in the limit $k \to \infty$,

$$\{\boldsymbol{h}^{\mathrm{ML}}(\boldsymbol{\sigma}),\boldsymbol{e}^{\mathrm{ML}}(\boldsymbol{\sigma},\boldsymbol{\omega})\} = \lim_{k\to\infty} \{\boldsymbol{h}^{(k)}(\boldsymbol{\sigma}),\boldsymbol{e}^{(k)}(\boldsymbol{\sigma},\boldsymbol{\omega})\}$$

or $\Delta h_i^{(k)}(\sigma,\omega) \to 0$ for $i = 1,\ldots, L$ and $\Delta e_{ij}^{(k)}(\sigma,\omega) \to 0$ for $1 \le i < j \le L$ and $\sigma,\omega \in \Omega \setminus \{\sigma_q\}$, a subset of $\Omega$ containing $q-1$ elements to account for gauge fixing.

## Pseudo-likelihood maximization

Besag [62] introduced the pseudo-likelihood as approximation to the likelihood function in which the global partition function is replaced by computationally tractable local estimates. The pseudo-likelihood inherits the concavity from the likelihood and yields the exact maximum-likelihood parameter in the limit of infinite data for Gaussian Markov random fields [41,62], but not in general [63]. Applications of this approximation to non-continuous categorical variables have been studied, for instance, in sparse inference of Ising models [64] but may lead to results that differ from the maximum-likelihood estimate. In this approach, the probability of the $m$-th observation, $\mathbf{x}^m$, is approximated by the product of the conditional probabilities of $x_r = x_r^m$ given observations in the remaining variables $\mathbf{x}_{\setminus r} := (x_1,\ldots,x_{r-1},x_{r+1},\ldots,x_L)^{\mathrm{T}} \in \Omega^{L-1}$ [51],

$$P(\mathbf{x}^m;\boldsymbol{h}(\boldsymbol{\sigma}),\boldsymbol{e}(\boldsymbol{\sigma},\boldsymbol{\omega})) \simeq \prod_{r=1}^{L} P(x_r = x_r^m | \mathbf{x}_{\setminus r} = \mathbf{x}_{\setminus r}^m;\boldsymbol{h}(\boldsymbol{\sigma}),\boldsymbol{e}(\boldsymbol{\sigma},\boldsymbol{\omega})).$$

Each factor is of the following analytical form,

$$P(x_r = x_r^m | \mathbf{x}_{\backslash r} = \mathbf{x}_{\backslash r}^m; \boldsymbol{h}(\boldsymbol{\sigma}), \boldsymbol{e}(\boldsymbol{\sigma}, \boldsymbol{\omega})) = \frac{\exp\left(h_r(x_r^m) + \sum_{j \neq r} e_{rj}(x_r^m, x_j^m)\right)}{\sum_{\sigma} \exp\left(h_r(\sigma) + \sum_{j \neq r} e_{rj}(\sigma, x_j^m)\right)},$$

which only depends on the unknown parameters $(e_{ij}(\sigma,\omega))_{i \neq r, j \neq r}$ and $(h_i(\sigma))_{i \neq r}$ and makes the computation of the pseudo-likelihood tractable. Note, we treat $e_{ij}(\sigma,\omega) = e_{ji}(\omega,\sigma)$ and $e_{ii}(\cdot,\cdot) = 0$. By this approximation, the loglikelihood Eq 21 becomes the pseudo-loglikelihood,

$$\ln l_{\mathrm{PL}}(\boldsymbol{h}(\boldsymbol{\sigma}), \boldsymbol{e}(\boldsymbol{\sigma}, \boldsymbol{\omega})) := \sum_{m=1}^{M} \sum_{r=1}^{L} \ln P(x_r = x_r^m | \mathbf{x}_{\backslash r} = \mathbf{x}_{\backslash r}^m; \boldsymbol{h}(\boldsymbol{\sigma}), \boldsymbol{e}(\boldsymbol{\sigma}, \boldsymbol{\omega})).$$

In the final formulation of the pseudo-likelihood maximization (PLM) problem, an $\ell^2$-regularizer is added to select for small absolute values of the inferred parameters,

$$\{\boldsymbol{h}^{\mathrm{PLM}}(\boldsymbol{\sigma}), \boldsymbol{e}^{\mathrm{PLM}}(\boldsymbol{\sigma}, \boldsymbol{\omega})\} = \underset{\boldsymbol{h}(\boldsymbol{\sigma}), \boldsymbol{e}(\boldsymbol{\sigma}, \boldsymbol{\omega})}{\arg\min} \{-\ln l_{\mathrm{PL}}(\boldsymbol{h}(\boldsymbol{\sigma}), \boldsymbol{e}(\boldsymbol{\sigma}, \boldsymbol{\omega})) + \lambda_{\boldsymbol{h}} \|\boldsymbol{h}(\boldsymbol{\sigma})\|_2^2 + \lambda_{\boldsymbol{e}} \|\boldsymbol{e}(\boldsymbol{\sigma}, \boldsymbol{\omega})\|_2^2\},$$

where $\lambda_{\boldsymbol{h}}, \lambda_{\boldsymbol{e}} > 0$ adjust the complexity of problem and are selected in a consistent manner across different protein families to avoid overfitting. This approach has been presented (with scaling of the pseudo-loglikelihood by $\frac{1}{M_{\mathrm{eff}}} w_m$ to include sequence weighting, see section "Sequence data preprocessing") by [51] under the name plmDCA (PseudoLikelihood Maximization Direct Coupling Analysis) and has shown performance improvements compared to the mean-field approximation Eq 20. Another inference method based on the pseudolikelihood maximization but including prior knowledge in terms of secondary structure and information on pairs likely to be in contact is Gremlin (Generative REgularized ModeLs of proteINs) [65–67].

## Sparse maximum likelihood

Similar to the derivation of the mean-field result (20), Jones et al. [8] approximated Eq 12 by a multivariate Gaussian and accessed the elements of the inverse covariance matrix by a maximum-likelihood inference under sparsity constraint [54,68,69]. The corresponding method has been called Psicov (Protein Sparse Inverse COVariance). The validity of this approach to solve the sparse maximum-likelihood problem in binary systems such as Ising models has been demonstrated by [69], followed by consistency studies [70]. In particular, the Psicov method infers the sparse maximum-likelihood estimate of the inverse covariance matrix Eq 19 for $\delta = 1$ using the analogue of the empirical covariance matrix derived from the observed amino acid frequencies, $\hat{C}(\boldsymbol{\sigma}, \boldsymbol{\omega})$. Its elements $\hat{C}_{ij}(\sigma, \omega) = f_{ij}(\sigma, \omega) - f_i(\sigma) f_j(\omega)$, the empirical connected correlations, are preprocessed by reweighting and regularized by pseudocounts and shrinkage. Regularized loglikelihood maximization Eq 19 selects a unique representation of the model, i.e., no additional gauge fixing is required. Using identity Eq 16 on the elements of the sparse maximum-likelihood (SML) estimate of the inverse covariance, $C_{1,\lambda}^{-1}(\boldsymbol{\sigma}, \boldsymbol{\omega})$, yields the estimates for the Lagrange multipliers,

$$\gamma_{ij}^{\mathrm{SML}}(\sigma, \omega) = -\frac{1}{2}(C_{1,\lambda}^{-1})_{ij}(\sigma, \omega) \quad \Rightarrow \quad e_{ij}^{\mathrm{SML}}(\sigma, \omega) = -(C_{1,\lambda}^{-1})_{ij}(\sigma, \omega)$$

for $\sigma,\omega \in \Omega$; in the second identity, the symmetric Lagrange multipliers $\gamma_{ij}(\sigma,\omega)$ defined for indices $i,j = 1,\ldots, L$ have been hypothetically translated to the reduced parameter formulation $e_{ij}(\sigma,\omega)$ for $1 \leq i < j \leq L$.

## Sequence data preprocessing

The study of residue–residue co-evolution is based on data from multiple sequence alignments, which represent sampling from the evolutionary record of a protein family. Multiple sequence alignments from currently existing sequence databases do not evenly represent the space of evolved sequences as they are subject to acquisition bias towards available species of interest. To account for uneven representation, sequence reweighting has been introduced to lower the contributions of highly similar sequences and assign higher weight to unique ones (see Durbin et al. [44], p. 124 ff.). In particular, the weight of the $m$-th sequence, $w_m := 1/k_m$, in the alignment $\{\mathbf{x}^1,\ldots,\mathbf{x}^M\}$, can be chosen to be the inverse of $k_m := \sum_{n=1}^{M} H\left(\sum_{i=1}^{L} \mathbf{1}(x_i^m, x_i^n) - L\cdot\theta\right)$, the number of sequences $\mathbf{x}^m$ shares more than $\theta \cdot 100\%$ of its residues with. Here, $\theta$ denotes a similarity threshold and is typically chosen as $0.7 \leq \theta \leq 0.9$, $\mathbf{1}(a,b) = 1$ if $a = b$ and $\mathbf{1}(a,b) = 0$, otherwise, and $H$ is the step function with $H(y) = 0$ if $y < 0$ and $H(y) = 1$, otherwise. This also provides us with an estimate of the effective number of sequences in the alignment, $M_{\text{eff}} := \sum_{m=1}^{M} w_m$. Additionally, pseudocount regularization with $\tilde{\lambda} > 0$ is used to deal with finite sampling bias and to account for underrepresentation [5–8,44,48], resulting in zero entries in $\hat{C}(\boldsymbol{\sigma},\boldsymbol{\omega})$, for instance, if a certain amino acid pair is never observed. The use of pseudocounts is equivalent to a maximum a posteriori (MAP) estimate under a specific inverse Wishart prior on the covariance matrix [48]. Both preprocessing steps combined yield the reweighted single and pair frequency counts,

$$f_i(\sigma) = \frac{1}{M_{\text{eff}} + \tilde{\lambda}}\left(\frac{\tilde{\lambda}}{q} + \sum_{m=1}^{M} w_m x_i^m(\sigma)\right), \qquad f_{ij}(\sigma,\omega) = \frac{1}{M_{\text{eff}} + \tilde{\lambda}}\left(\frac{\tilde{\lambda}}{q^2} + \sum_{m=1}^{M} w_m x_i^m(\sigma) x_j^m(\omega)\right),$$

in residues $i,j = 1,\ldots, L$ and for amino acids $\sigma,\omega\in\Omega$. Ideally for maximum-likelihood inference, the random variables are assumed to be independent and identically distributed. However, this is typically violated in realistic sequence data due to phylogenetic and sequencing bias, and the reweighting presented here does not necessarily solve this problem.

# Scoring Functions for the Pairwise Interaction Strengths

For pairwise maximum-entropy models of continuous variables, the natural scoring function for the interaction strength between two variables $x_i$ and $x_j$, given the inferred inverse covariance matrix, is the partial correlation Eq 18. However, for categorical variables, the situation is more complicated, and there are several alternative choices of scoring functions. Requirements on the scoring function are that it has to account for the chosen gauge and, in the case of protein contact prediction, evaluate the coupling strength between two residues $i$ and $j$ summarized across all possible $q^2$ amino acids pairs. The highest scoring residue pair is, for instance, used to predict the 3-D structure of the protein of interest. For this purpose, the direct information, defined as the mutual information applied to $P_{ij}^{\text{dir}}(\sigma,\omega) = \frac{1}{Z_{ij}}\exp(e_{ij}(\sigma,\omega) + \tilde{h}_i(\sigma) + \tilde{h}_j(\omega))$ instead of $f_{ij}(\sigma,\omega)$,

$$\mathrm{DI}_{ij} = \sum_{\sigma,\omega\in\Omega} P_{ij}^{\text{dir}}(\sigma,\omega) \ln\left(\frac{P_{ij}^{\text{dir}}(\sigma,\omega)}{f_i(\sigma) f_j(\omega)}\right),$$

has been introduced [5]. In $P_{ij}^{\text{dir}}(\sigma,\omega)$, $\tilde{h}_i(\sigma)$ and $\tilde{h}_j(\omega)$ are chosen to be consistent with the (reweighted and regularized) single-site frequency counts, $f_i(\sigma)$ and $f_j(\omega)$, and $Z_{ij}$ such that the sum over all pairs $(i, j)$ with $1 \leq i < j \leq L$ is normalized to 1. The direct information is invariant under gauge changes of the Hamiltonian $\mathcal{H}$, which means that any suitable gauge choice results in the same scoring values. As an alternative measure of the interaction strength for a particular pair $(i, j)$, the Frobenius norm of the 21×21-submatrices of $(e_{ij}(\sigma,\omega))_{\sigma,\omega}$ has been used,

$$\|e_{ij}\|_{\mathrm{F}} = \left(\sum_{\sigma,\omega\in\Omega} e_{ij}(\sigma,\omega)^2\right)^{1/2}.$$

However, this expression is not gauge-invariant [5]. In this context, the notation with $e_{ij}(\sigma, \omega)$, which refers to indices restricted to $i < j$, is extended and treated such that $e_{ij}(\sigma,\omega) = e_{ji}(\omega, \sigma)$ and $e_{ij}(\cdot,\cdot) = 0$; then $||e_{ij}||_{\mathrm{F}} = ||e_{ji}||_{\mathrm{F}}$ and $||e_{ii}||_{\mathrm{F}} = 0$. In order to correct for phylogenetic biases in the identification of co-evolved residues, Dunn et al. [27] introduced the average product correction (APC). It has been originally used in combination with the mutual information but was recently combined with the $\ell^1$-norm [8] and the Frobenius/$\ell^2$-norm [51] and is derived from the averages over rows and columns of the corresponding norm of the matrix of the $e_{ij}$ parameters. In this formulation, the pair scoring function is

$$\text{APC-FN}_{ij} = \|e_{ij}\|_{\mathrm{F}} - \frac{\|e_{i\cdot}\|_{\mathrm{F}}\|e_{\cdot j}\|_{\mathrm{F}}}{\|e_{\cdot\cdot}\|_{\mathrm{F}}} \tag{24}$$

for $e_{ij}$-parameters fixed by zero-sum gauge and with the means over the non-zero elements in row, column and full matrix, $\|e_{i\cdot}\|_{\mathrm{F}} := \frac{1}{L-1}\sum_{j=1}^{L} \|e_{ij}\|_{\mathrm{F}}$, $\|e_{\cdot j}\|_{\mathrm{F}} := \frac{1}{L-1}\sum_{i=1}^{L} \|e_{ij}\|_{\mathrm{F}}$ and $\|e_{\cdot\cdot}\|_{\mathrm{F}} := \frac{1}{L(L-1)}\sum_{i,j=1}^{L} \|e_{ij}\|_{\mathrm{F}}$, respectively. Alternatively, the average product-corrected $\ell^1$-norm applied to the 20×20-submatrices of the estimated inverse covariance matrix, in which contributions from gaps are ignored, has been introduced by the authors of [8] as the Psicov-score. Using the average product correction, the authors of [51] showed for interaction parameters inferred by the mean-field approximation that scoring with the average product-corrected Frobenius norm increased the precision of the predicted contacts compared to scoring with the DI-score. The practical consequence of the choice of scoring method depends on the dataset and the parameter inference method.

## Discussion of Results, Improvements, and Applications

Maximum entropy-based inference methods can help in estimating interactions underlying biological data. This class of models, combined with suitable methods for inferring their numerical parameters, has been shown to reveal—to a reasonable approximation—the direct interactions in many biological applications, such as gene expression or protein residue—residue coevolution studies. In this review, we have presented maximum-entropy models for the continuous and categorical random variable case. Both approaches can be integrated into a framework, which allows the use of solutions obtained for continuous variables as approximations for the categorical random variable case (Fig 3).

The validity and precision of the available maximum-entropy methods could be improved to yield more biologically insightful results in several ways. Advanced approximation methods derived from Ising model approaches [59,71] are possible extensions for efficient inference. Moreover, additional terms beyond pair interactions can be included in models of continuous and discrete random variables [1,33,59]. However, higher-order models demand more data, which is a major bottleneck for their application to biological problems. In the case of protein contact prediction, this could be resolved by getting more sequences, which are being obtained

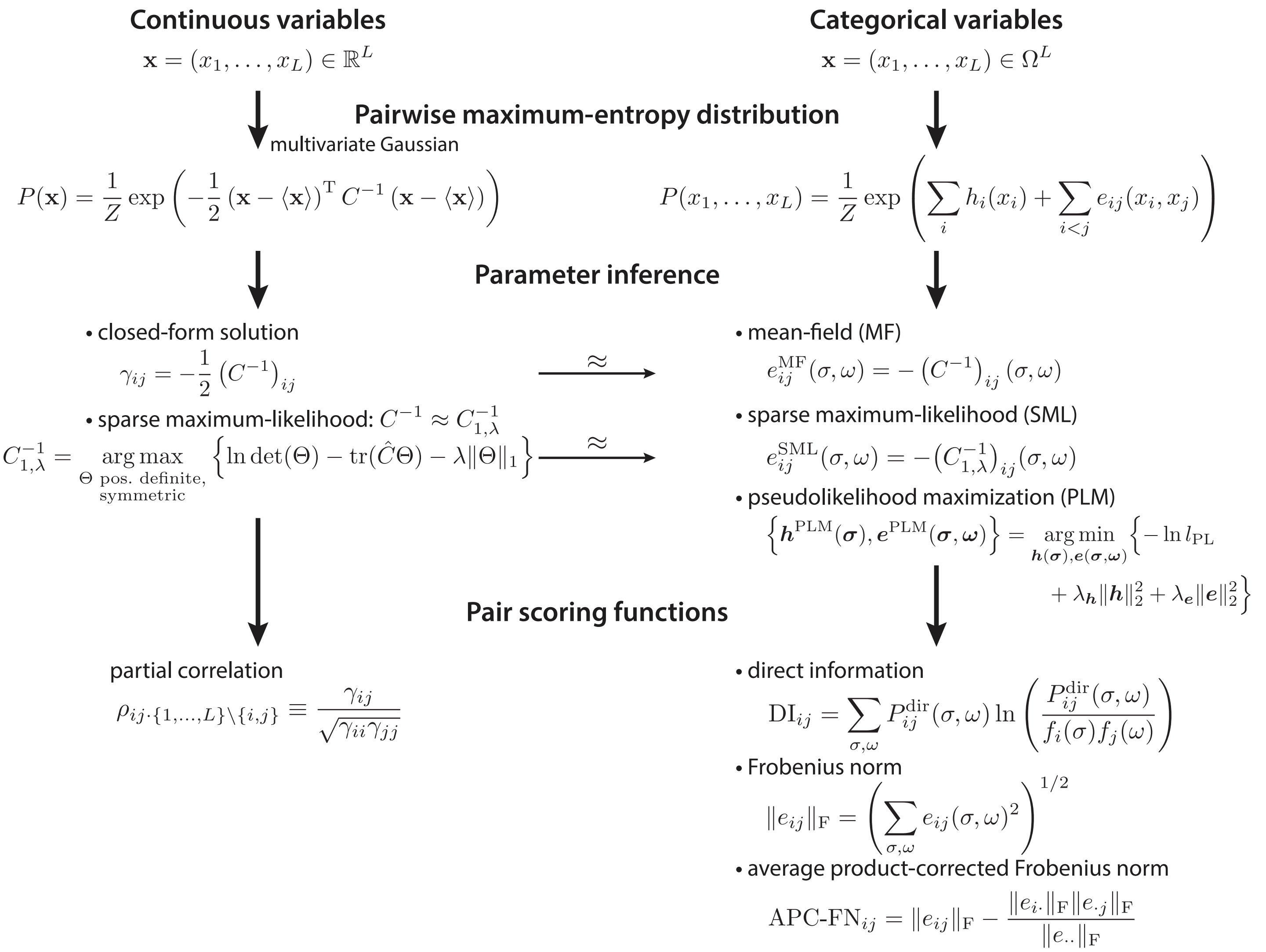
Continuous variables
$\mathbf{x} = (x_1, \ldots, x_L) \in \mathbb{R}^L$
Categorical variables
$\mathbf{x} = (x_1, \ldots, x_L) \in \Omega^L$
Pairwise maximum-entropy distribution
multivariate Gaussian
$P(\mathbf{x}) = \frac{1}{Z} \exp\left(-\frac{1}{2}(\mathbf{x} - \langle \mathbf{x} \rangle)^{\mathrm{T}} C^{-1} (\mathbf{x} - \langle \mathbf{x} \rangle)\right)$
$P(x_1, \ldots, x_L) = \frac{1}{Z} \exp\left(\sum_i h_i(x_i) + \sum_{i<j} e_{ij}(x_i, x_j)\right)$
Parameter inference
• closed-form solution
$\gamma_{ij} = -\frac{1}{2}\left(C^{-1}\right)_{ij}$
• sparse maximum-likelihood: $C^{-1} \approx C^{-1}_{1,\lambda}$
$C^{-1}_{1,\lambda} = \underset{\Theta \text{ pos. definite, symmetric}}{\arg\max} \left\{\ln\det(\Theta) - \mathrm{tr}(\hat{C}\Theta) - \lambda \|\Theta\|_1\right\}$
$\xrightarrow{\approx}$
$\xrightarrow{\approx}$
• mean-field (MF)
$e^{\mathrm{MF}}_{ij}(\sigma, \omega) = -\left(C^{-1}\right)_{ij}(\sigma, \omega)$
• sparse maximum-likelihood (SML)
$e^{\mathrm{SML}}_{ij}(\sigma, \omega) = -\left(C^{-1}_{1,\lambda}\right)_{ij}(\sigma, \omega)$
• pseudolikelihood maximization (PLM)
$\left\{\boldsymbol{h}^{\mathrm{PLM}}(\boldsymbol{\sigma}), \boldsymbol{e}^{\mathrm{PLM}}(\boldsymbol{\sigma}, \boldsymbol{\omega})\right\} = \underset{\boldsymbol{h}(\boldsymbol{\sigma}), \boldsymbol{e}(\boldsymbol{\sigma}, \boldsymbol{\omega})}{\arg\min} \left\{-\ln l_{\mathrm{PL}} + \lambda_{\boldsymbol{h}} \|\boldsymbol{h}\|_2^2 + \lambda_{\boldsymbol{e}} \|\boldsymbol{e}\|_2^2\right\}$
Pair scoring functions
partial correlation
$\rho_{ij \cdot \{1,\ldots,L\} \setminus \{i,j\}} \equiv \frac{\gamma_{ij}}{\sqrt{\gamma_{ii}\gamma_{jj}}}$
• direct information
$\mathrm{DI}_{ij} = \sum_{\sigma,\omega} P^{\mathrm{dir}}_{ij}(\sigma, \omega) \ln\left(\frac{P^{\mathrm{dir}}_{ij}(\sigma, \omega)}{f_i(\sigma) f_j(\omega)}\right)$
• Frobenius norm
$\|e_{ij}\|_{\mathrm{F}} = \left(\sum_{\sigma,\omega} e_{ij}(\sigma, \omega)^2\right)^{1/2}$
• average product-corrected Frobenius norm
$\text{APC-FN}_{ij} = \|e_{ij}\|_{\mathrm{F}} - \frac{\|e_{i\cdot}\|_{\mathrm{F}} \|e_{\cdot j}\|_{\mathrm{F}}}{\|e_{\cdot\cdot}\|_{\mathrm{F}}}$

**Fig 3. Scheme of pairwise maximum-entropy probability models.** The maximum-entropy probability distribution with pairwise constraints for continuous random variables is the multivariate Gaussian distribution (left column). For the maximum-entropy probability distribution in the categorical variable case (right column), various approximative solutions exist, e.g., the mean-field, the sparse maximum-likelihood, and the pseudolikelihood maximization solution. The mean-field and the sparse maximum-likelihood result can be derived from the Gaussian approximation of binarized categorical variables (thin arrow). Pair scoring functions for the continuous case are the partial correlations (left column). For the categorical variable case, the direct information, the Frobenius norm, and the average product-corrected Frobenius norm are used to score pair couplings from the inferred parameters (right column).

doi:10.1371/journal.pcbi.1004182.g003

as the result of extraordinary advances in sequencing technology. The quality of existing methods can be improved by careful refinement of sequence alignments in terms of cutoffs and gaps or by attaching optimized weights to each of the data sequences. Alternatively, one could try to improve the existing model frameworks by accounting for phylogenetic progression [27,49,72] and finite sampling biases.

The advancement of inference methods for biological datasets could help solve many interesting biological problems, such as protein design or the analysis of multi-gene effects in relating variants to phenotypic changes as well as multi-genic traits [73,74]. The methods presented here could help reduce the parameter space of genome-wide association studies to first approximation. In particular, we envision the following applications: (1) in the disease context, co-evolution studies of oncogenic events, for example copy number alterations, mutations, fusions

**Table 1. Overview of software tools to infer pairwise interactions from datasets in continuous or categorical variables with maximum-entropy/GGM-based methods.**

| Data type | Method | Name | Output | Link |
|---|---|---|---|---|
| categorical | mean-field | DCA, mfDCA | DI-score | [76,77] |
| | pseudolikelihood maximization | plmDCA | APC-FN-score | [78–80] |
| | pseudolikelihood maximization | Gremlin | Gremlin-score | [81] |
| | sparse maximum-likelihood | Psicov | Psicov-score | [82] |
| continuous | sparse maximum-likelihood | glasso | partial correlations | [83] |
| | $\ell^2$-regularized maximum-likelihood | scout | partial correlations | [84] |
| | shrinkage | corpcor, GeneNet | partial correlations | [85,86] |

doi:10.1371/journal.pcbi.1004182.t001

and alternative splicing, can be used to derive direct co-evolution signatures of cancer from available data, such as The Cancer Genome Atlas (TCGA); (2) *de novo* design of protein sequences as, for example, described in [65,75] for the WW domain using design rules based on the evolutionary information extracted from the multiple sequence alignment; and (3) develop quantitative models of protein fitness computed from sequence information.

In general, in a complex biological system, it is often useful for descriptive and predictive purposes to derive the interactions that define the properties of the system. With the methods presented here and available software (Table 1), our goal is not only to describe how to infer these interactions but also to highlight tools for the prediction and redesign of properties of biological systems.

## Acknowledgments

We thank Theofanis Karaletsos, Sikander Hayat, Stephanie Hyland, Quaid Morris, Deb Bemis, Linus Schumacher, John Ingraham, Arman Aksoy, Julia Vogt, Thomas Hopf, Andrea Pagnani, and Torsten Groß for insightful discussions.

## References

1. Lezon TR, Banavar JR, Cieplak M, Maritan A, Fedoroff NV. Using the principle of entropy maximization to infer genetic interaction networks from gene expression patterns. Proceedings of the National Academy of Sciences of the United States of America. 2006; 103(50):19033–19038. PMID: 17138668
2. Locasale JW, Wolf-Yadlin A. Maximum entropy reconstructions of dynamic signaling networks from quantitative proteomics data. PloS one. 2009; 4(8):e6522. doi: 10.1371/journal.pone.0006522 PMID: 19707567
3. Schneidman E, Berry II MJ, Segev R, Bialek W. Weak pairwise correlations imply strongly correlated network states in a neural population. Nature. 2006; 440:1007–1012. PMID: 16625187
4. Tang A, Jackson D, Hobbs J, Chen W, Smith JL, Patel H, et al. A maximum entropy model applied to spatial and temporal correlations from cortical networks in vitro. The Journal of Neuroscience. 2008; 28 (2):505–518. doi: 10.1523/JNEUROSCI.3359-07.2008 PMID: 18184793
5. Weigt M, White RA, Szurmant H, Hoch JA, Hwa T. Identification of direct residue contacts in protein—protein interaction by message passing. Proceedings of the National Academy of Sciences of the United States of America. 2009; 106(1):67–72. doi: 10.1073/pnas.0805923106 PMID: 19116270
6. Marks DS, Colwell LJ, Sheridan R, Hopf TA, Pagnani A, Zecchina R, et al. Protein 3D Structure Computed from Evolutionary Sequence Variation. PLoS One. 2011; 6(12):e28766. doi: 10.1371/journal.pone.0028766 PMID: 22163331
7. Morcos F, Pagnani A, Lunt B, Bertolino A, Marks D, Sander C, et al. Direct-coupling analysis of residue co-evolution captures native contacts across many protein families. Proceedings of the National Academy of Sciences of the United States of America. 2011; 108:E1293–E1301. doi: 10.1073/pnas.1111471108 PMID: 22106262
8. Jones DT, Buchan DWA, Cozzetto D, Pontil M. PSICOV: precise structural contact prediction using sparse inverse covariance estimation on large multiple sequence alignments. Bioinformatics. 2012; 28 (2):184–190. doi: 10.1093/bioinformatics/btr638 PMID: 22101153

9. Stephens GJ, Bialek W. Statistical mechanics of letters in words. Physical Review E. 2010; 81 (6):066119.

10. Bialek W, Cavagna A, Giardina I, Mora T, Silvestri E, Viale M, et al. Statistical mechanics for natural flocks of birds. Proceedings of the National Academy of Sciences. 2012; 109(13):4786–4791.

11. Wood K, Nishida S, Sontag ED, Cluzel P. Mechanism-independent method for predicting response to multidrug combinations in bacteria. Proceedings of the National Academy of Sciences of the United States of America. 2012; 109(30):12254–12259. doi: 10.1073/pnas.1201281109 PMID: 22773816

12. Whittaker J. Graphical models in applied multivariate statistics. Wiley Publishing; 2009.

13. Lauritzen SL. Graphical models. Oxford: Oxford University Press; 1996.

14. Butte AJ, Kohane IS. Unsupervised knowledge discovery in medical databases using relevance networks. In: Proceedings of the AMIA Symposium. American Medical Informatics Association; 1999. p. 711–715.

15. Toh H, Horimoto K. Inference of a genetic network by a combined approach of cluster analysis and graphical Gaussian modeling. Bioinformatics. 2002; 18(2):287–297. PMID: 11847076

16. Dobra A, Hans C, Jones B, Nevins JR, Yao G, West M. Sparse graphical models for exploring gene expression data. Journal of Multivariate Analysis. 2004; 90(1):196–212.

17. Schäfer J, Strimmer K. A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics. Statistical applications in genetics and molecular biology. 2005; 4(1):1–32.

18. Roudi Y, Nirenberg S, Latham PE. Pairwise maximum entropy models for studying large biological systems: when they can work and when they can't. PLoS Computational Biology. 2009; 5(5):e1000380. doi: 10.1371/journal.pcbi.1000380 PMID: 19424487

19. Cramér H. Mathematical methods of statistics. vol. 9. Princeton university press; 1999.

20. Guttman L. A note on the derivation of formulae for multiple and partial correlation. The Annals of Mathematical Statistics. 1938; 9(4):305–308.

21. Krumsiek J, Suhre K, Illig T, Adamski J, Theis FJ. Gaussian graphical modeling reconstructs pathway reactions from high-throughput metabolomics data. BMC Systems Biology. 2011; 5(1):21.

22. Giraud BG and Heumann, John M and Lapedes, Alan S. Superadditive correlation. Physical Review E. 1999; 59(5):4983–4991.

23. Neher E. How frequent are correlated changes in families of protein sequences? Proceedings of the National Academy of Sciences of the United States of America. 1994; 91(1):98–102. PMID: 8278414

24. Göbel U, Sander C, Schneider R, Valencia A. Correlated mutations and residue contacts in proteins. Proteins. 1994; 18(4):309–317. PMID: 8208723

25. Taylor WR, Hatrick K. Compensating changes in protein multiple sequence alignments. Protein Engineering. 1994; 7(3):341–348. PMID: 8177883

26. Shindyalov IN and Kolchanov NA and Sander C. Can three-dimensional contacts in protein structures be predicted by analysis of correlated mutations? Protein Engineering. 1994; 7(3):349–358. PMID: 8177884

27. Dunn SD, Wahl LM, Gloor GB. Mutual information without the influence of phylogeny or entropy dramatically improves residue contact prediction. Bioinformatics. 2008; 24(3):333–340. PMID: 18057019

28. Burger L, Van Nimwegen E. Disentangling direct from indirect co-evolution of residues in protein alignments. PLoS computational biology. 2010; 6(1):e1000633. doi: 10.1371/journal.pcbi.1000633 PMID: 20052271

29. Atchley WR, Wollenberg KR, Fitch WM, Terhalle W, Dress AW. Correlations among amino acid sites in bHLH protein domains: an information theoretic analysis. Molecular biology and evolution. 2000; 17 (1):164–178. PMID: 10666716

30. Lapedes A, Giraud B, Jarzynski C. Using Sequence Alignments to Predict Protein Structure and Stability With High Accuracy. eprint arXiv:12072484. 2002;.

31. Ciriello G, Miller ML, Aksoy BA, Senbabaoglu Y, Schultz N, Sander C. Emerging landscape of oncogenic signatures across human cancers. Nature genetics. 2013; 45(10):1127–1133. doi: 10.1038/ng.2762 PMID: 24071851

32. Koller D, Friedman N. Probabilistic graphical models: principles and techniques. MIT press; 2009.

33. Mead LR, Papanicolaou N. Maximum entropy in the problem of moments. Journal of Mathematical Physics. 1984; 25:2404–2417.

34. MacKay DJ. Information theory, inference and learning algorithms. Cambridge university press; 2003.

35. Cover TM, Thomas AJ. Elements of information theory. John Wiley & Sons; 2012.

36. Agmon N, Alhassid Y, Levine RD. An algorithm for finding the distribution of maximal entropy. Journal of Computational Physics. 1979; 30(2):250–258.

37. Shannon CE. A Mathematical Theory of Communication. Bell system technical journal. 1948; 27 (3):379–423.

38. Jaynes ET. Information Theory and Statistical Mechanics. Physical Review. 1957; 106(4):620–630.

39. Jaynes ET. Information Theory and Statistical Mechanics II. Physical Review. 1957; 108(2):171–190.

40. Jaynes ET. Probability theory: the logic of science. Cambridge: Cambridge university press; 2003.

41. Murphy KP. Machine learning: a probabilistic perspective. The MIT Press; 2012.

42. Balescu R. Matter out of Equilibrium. World Scientific; 1997.

43. Goldstein S, Lebowitz JL. On the (Boltzmann) entropy of non-equilibrium systems. Physica D: Nonlinear Phenomena. 2004; 193(1):53–66.

44. Durbin R, Eddy SR, Krogh A, Mitchison G. Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids. Cambridge University Press; 1998.

45. Finn RD, Bateman A, Clements J, Coggill P, Eberhardt RY, Eddy SR, et al. Pfam: the protein families database. Nucleic acids research. 2014; 42:D222–D230. doi: 10.1093/nar/gkt1223 PMID: 24288371

46. Remmert M, Biegert A, Hauser A, Söding J. HHblits: lightning-fast iterative protein sequence searching by HMM-HMM alignment. Nature methods. 2012; 9(2):173–175.

47. Finn RD, Clements J, Eddy SR. HMMER web server: interactive sequence similarity searching. Nucleic acids research. 2011;p. gkr367.

48. Baldassi C, Zamparo M, Feinauer C, Procaccini A, Zecchina R, Weigt M, et al. Fast and accurate multivariate Gaussian modeling of protein families: Predicting residue contacts and protein-interaction partners. PloS one. 2014; 9(3):e92721. doi: 10.1371/journal.pone.0092721 PMID: 24663061

49. Lapedes AS, Giraud BG, Liu LC, Stormo GD. A Maximum Entropy Formalism for Disentangling Chains of Correlated Sequence Positions. In: Proceedings of the IMS/AMS International Conference on Statistics in Molecular Biology and Genetics; 1998. p. 236–256.

50. Santolini M, Mora T, Hakim V. A general pairwise interaction model provides an accurate description of in vivo transcription factor binding sites. PloS one. 2014; 9(6):e99015. doi: 10.1371/journal.pone.0099015 PMID: 24926895

51. Ekeberg M, Lövkvist C, Lan Y, Weigt M, Aurell E. Improved contact prediction in proteins: Using pseudolikelihoods to infer Potts models. Physical Review E. 2013; 87(1):012707.

52. Bishop CM. Pattern recognition and machine learning. New York: Springer-Verlag; 2006.

53. Meinshausen N, Bühlmann P. High-dimensional graphs and variable selection with the lasso. The Annals of Statistics. 2006; 34(3):1436–1462.

54. Friedman J, Hastie T, Tibshirani R. Sparse inverse covariance estimation with the graphical lasso. Biostatistics. 2008; 9(3):432–441. PMID: 18079126

55. Witten DM, Tibshirani R. Covariance-regularized regression and classification for high dimensional problems. Journal of the Royal Statistical Society: Series B (Statistical Methodology). 2009; 71(3):615–636.

56. Ledoit O, Wolf M. A well-conditioned estimator for large-dimensional covariance matrices. Journal of multivariate analysis. 2004; 88(2):365–411.

57. Kappen HJ, Rodriguez F. Efficient learning in Boltzmann machines using linear response theory. Neural Computation. 1998; 10(5):1137–1156.

58. Tanaka T. Mean-field theory of Boltzmann machine learning. Physical Review E. 1998; 58(2):2302–2310.

59. Roudi Y, Aurell E, Hertz JA. Statistical physics of pairwise probability models. Frontiers in computational neuroscience. 2009;3. doi: 10.3389/neuro.10.003.2009 PMID: 19242556

60. Wainwright MJ, Jordan MI. Graphical models, exponential families, and variational inference. Foundations and Trends in Machine Learning. 2008; 1(1–2):1–305.

61. Broderick T, Dudik M, Tkacik G, Schapire RE, Bialek W. Faster solutions of the inverse pairwise Ising problem. arXiv preprint arXiv:07122437. 2007;.

62. Besag J. Statistical analysis of non-lattice data. The Statistician. 1975; 24(3):179–195.

63. Liang P, Jordan MI. An asymptotic analysis of generative, discriminative, and pseudolikelihood estimators. In: Proceedings of the 25th international conference on Machine learning. ACM; 2008. p. 584–591.

64. Höfling H, Tibshirani R. Estimation of sparse binary pairwise markov networks using pseudo-likelihoods. The Journal of Machine Learning Research. 2009; 10:883–906.

65. Balakrishnan S, Kamisetty H, Carbonell JG, Lee SI, Langmead CJ. Learning generative models for protein fold families. Proteins: Structure, Function, and Bioinformatics. 2011; 79(4):1061–1078.

66. Kamisetty H, Ovchinnikov S, Baker D. Assessing the utility of coevolution-based residue—residue contact predictions in a sequence-and structure-rich era. Proceedings of the National Academy of Sciences. 2013; 110(39):15674–15679.

67. Ovchinnikov S, Kamisetty H, Baker D. Robust and accurate prediction of residue—residue interactions across protein interfaces using evolutionary information. eLife. 2014; 3: e02030. doi: 10.7554/eLife.02030 PMID: 24842992

68. Wainwright MJ, Jordan MI. Log-determinant relaxation for approximate inference in discrete Markov random fields. IEEE Transactions on Signal Processing. 2006; 54(6):2099–2109.

69. Banerjee O, El Ghaoui L, d'Aspremont A. Model selection through sparse maximum likelihood estimation for multivariate gaussian or binary data. The Journal of Machine Learning Research. 2008; 9:485–516.

70. Ravikumar P, Wainwright MJ, Lafferty JD. High-dimensional Ising model selection using l1-regularized logistic regression. The Annals of Statistics. 2010; 38(3):1287–1319.

71. Sessak V, Monasson R. Small-correlation expansions for the inverse Ising problem. Journal of Physics A: Mathematical and Theoretical. 2009; 42(5):055001.

72. Lapedes AS, Giraud BG, Liu L, Stormo GD. Correlated mutations in models of protein sequences: phylogenetic and structural effects. In: Statistics in Molecular Biology/IMS Lecture Notes—Monograph Series. JSTOR; 1999. p. 236–256.

73. Rockman MV. Reverse engineering the genotype—phenotype map with natural genetic variation. Nature. 2008; 456(7223):738–744. doi: 10.1038/nature07633 PMID: 19079051

74. Ritchie MD, Holzinger ER, Li R, Pendergrass SA, Kim D. Methods of integrating data to uncover genotype-phenotype interactions. Nature Reviews Genetics. 2015; 16(2):85–97. doi: 10.1038/nrg3868 PMID: 25582081

75. Russ WP, Lowery DM, Mishra P, Yaffe MB, Ranganathan R. Natural-like function in artificial WW domains. Nature. 2005; 437(7058):579–583. PMID: 16177795

76. EVFold. http://evfold.org.

77. Direct Coupling Analysis. http://dca.rice.edu.

78. Ekeberg M. pseudolikelihood maximization Direct-Coupling Analysis. http://plmdca.csc.kth.se.

79. Pagnani A. Pseudo Likelihood Maximization for protein in Julia. https://github.com/pagnani/PlmDCA.

80. CCMpred. https://bitbucket.org/soedinglab/ccmpred.

81. Gremlin. http://gremlin.bakerlab.org.

82. Psicov. http://bioinfadmin.cs.ucl.ac.uk/downloads/PSICOV.

83. Friedman J, Hastie T, Tibshirani R. Graphical lasso in R and Matlab. http://statweb.stanford.edu/~tibs/glasso/.

84. Witten DM, Tibshirani R. scout: Implements the Scout method for Covariance-Regularized Regression. http://cran.r-project.org/web/packages/scout/index.html.

85. Schäfer J, Opgen-Rhein R, Strimmer K. Modeling and Inferring Gene Networks. http://strimmerlab.org/software/genenet/.

86. Schäfer J, Opgen-Rhein R, Zuber V, Ahdesmäki M, Silva APD, Strimmer K. Efficient Estimation of Covariance and (Partial) Correlation. http://strimmerlab.org/software/corpcor/.