Inference of compressed Potts graphical models

Francesca Rizzato, Alice Coucke, Eleonora de Leonardis, J. P. P. Barton, Jérôme Tubiana, Remi Monasson, Simona Cocco

Inference of compressed Potts graphical models

Francesca Rizzato[a],[1] Alice Coucke[b],[1] Eleonora de Leonardis[c],[1] John P. Barton,[2] Jérôme Tubiana[d],[1] Rémi Monasson,[1] and Simona Cocco[1]

[1] *Laboratoire de Physique, Ecole Normale Supérieure and CNRS-UMR8023, PSL Research University, 24 Rue Lhomond, 75005 Paris, France*
[2] *Department of Physics and Astronomy, University of California, Riverside, 900 University Ave, Riverside, CA, United States*

We consider the problem of inferring a graphical Potts model on a population of variables. This inverse Potts problem generally involves the inference of a large number of parameters, often larger than the number of available data, and, hence, requires the introduction of regularization. We study here a double regularization scheme, in which the number of Potts states (colors) available to each variable is reduced, and interaction networks are made sparse. To achieve the color compression, only Potts states with large empirical frequency (exceeding some threshold) are explicitly modeled on each site, while the others are grouped into a single state. We benchmark the performances of this mixed regularization approach, with two inference algorithms, Adaptive Cluster Expansion (ACE) and PseudoLikelihood Maximization (PLM) on synthetic data obtained by sampling disordered Potts models on an Erdős-Rényi random graphs. We show in particular that color compression does not affect the quality of reconstruction of the parameters corresponding to high-frequency symbols, while drastically reducing the number of the other parameters and thus the computational time. Our procedure is also applied to multi-sequence alignments of protein families, with similar results.

I. INTRODUCTION

Graphical models are an important tool to model dependencies and infer effective interactions among a population of variables from data. We hereafter refer to this approach as the inverse Ising problem [1–3] in the case of binary variables, and as the inverse Potts problem in the more general case of of multi-categorial variables[4]. Applications include inferring functional couplings among a set of neurons from the recording of their activity [5–7], between birds in flocks [8] and among amino acids in sequences that belong to the same protein family [9]. Over the last decade, it was shown that describing these protein families by Potts models, whose parameters were learned from the corresponding sequence alignment could provide information on the protein structure [9–18], help predict fitness variations following mutations [3, 19–22] and design new working proteins of the same family [23–25]. Given the computational untractability of achieving exact solutions, different effective methods have been proposed to infer the Potts parameters from sequence data, including Gaussian approximation with different priors [10, 17, 26, 27], message passing [11], PseudoLikelihood Maximization (PLM)[9, 28, 29], minimum probability flow [30], and the Adaptive Cluster Expansion (ACE) method [31, 32].

Even if Potts models only include pairwise interactions, the number of parameters to be inferred is huge. For an N-site protein, where each site can be one out of the 20 natural amino acids or an extra symbol standing for a site insertion or deletion, the number of independent parameters is $20 \cdot N + 20^2 \cdot N(N-1)/2$. This gives about 10^6 parameters for $N = 100$ and almost 10^8 parameters for $N = 500$, while protein sequence alignments typically include few thousands to few tens thousand sequences. Moreover, amino-acid frequencies, and, hence, sampling quality may vary substantially from site to site, making it impossible to accurately reconstruct the complete set of Potts parameters. To overcome the problem of undersampling regularization terms are generally included. Standard L_2-regularization helps constraining parameter values, but does not change their number. L_1 and L_0-based regularization, on the contrary, may effectively remove many interaction parameters associated to low (in absolute value) connected correlations.

In this paper a simple procedure to reduce the number q of Potts parameters is described and analyzed. In physics Potts states are often referred to as colors, so we call this state reduction procedure *color compression*. Our goal is to infer a compressed Potts model, where the number of states $q_i \leq q$ depends on the site i. The basic idea is to group together rarely observed states on each site, defined as those below a given

[a] Now at SISSA Medialab Via Bonomea, 265, 34136 Trieste, Italy
[b] Now at Snips, 18 rue Saint Marc, 75002 Paris, France
[c] Now at Groupe Galeries Lafayette, 44 rue de Chateaudun, 75009 Paris, France
[d] Now at Blavatnik School of Computer Science, Tel Aviv University, Israel

frequency threshold f_0. This way, the number of Potts states q_i on each site i is variable, leading to the reduced number of parameters $M_{cc} = \sum_i^N q_i + \sum_{i<j}^N q_i\, q_j$. Color compression is therefore equivalent to an L_0 regularization on the number of inferred parameters associated to the Potts states. Slightly different schemes are based on grouping colors according to their entropy contributions to the site variability [32] or to their mutual information [33] or compression to a fixed number of colors [34], and comes to a similar outcome. As expected color compression can help in limiting the computational time, in avoiding overfitting and, in a more theoretical framework, in understanding the intrinsic dimensionality of the problem, by distinguishing the parameters that can be reliably inferred from those fixed through regularization only. Color compression was already used by some of us in the Adaptive Cluster Expansion (ACE) algorithm [32] in order to reduce the computational time and have a simpler inferred model for the analysis of protein sequence data, but its performance has not been systematically tested up to now.

The aim of the present study is to benchmark the color compression procedure in a systematic way as a function of the compression frequency threshold f_0 and of the sampling depth B: Correlations between variables that are grouped in the color compression procedure are lost. We expect therefore that compression is useful in the inference at low B and low f_0 to discard correlations between poorly sampled colors, affected by large statistical fluctuations. Conversely, information on the correlated structure of the data will be lost for large B and f_0. Consequences of compression on the performances of inferred models are here investigated for different regularization types and strengths, for two particular inference procedures: PLM with large or small L_2 regularization parameters, and the ACE procedure with a fully connected or sparse interaction graph. We then introduce a procedure to recover, after inference of the compressed Potts model, a full Potts model over all possible q states, which we refer to as color decompression. Decompression is necessary to compare the inferred parameters to the ground truth (when available), to compare the quality of the inferences for different color compression strengths (and therefore different q_i values), and, more generally, when the model is used to predict the behavior of poorly sampled variables in the original data set.

The first part of the paper briefly sketches the methodological background and the inference algorithms (Section II). In Section III the procedures of color compression and decompression are introduced. We then assess the performances of the procedures on synthetic data generated from Potts model on random graphs in Section IV. In Section V we show an illustrative example on fitness prediction for real proteins, to verify that the results obtained on synthetic data model translate to real cases. Section VI shows the gain our color compression procedure provides in terms of computational time. Some conclusion and perspectives are presented in Sec. VII.

II. REMINDER ON INFERENCE AND ALGORITHMS

A. Inverse Potts Problem

The Potts model describes a system of N interacting sites, each assuming one of q possible Potts states (or colors). The probability distribution of each color on each site is controlled by a set of parameters that can be divided into local fields $h_i(a_i)$, depending only on one site i and its color a_i, and pairwise couplings $J_{ij}(a_i, a_j)$, depending on the pair of sites i, j and the two Potts states a_i, a_j. An energy value is associated to each system configuration $\mathbf{a} = a_1, \dots a_N$,

$$E(\mathbf{a}|\mathbf{J}) = -\sum_{i=1}^{N} h_i(a_i) - \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} J_{ij}(a_i, a_j) \tag{1}$$

and, consequently, a probability

$$P(\mathbf{a}|\mathbf{J}) = \frac{\exp\left(-E(\mathbf{a}|\mathbf{J})\right)}{Z(\mathbf{J})}\,, \tag{2}$$

where $Z(\mathbf{J}) = \sum_{\mathbf{a}} \exp\left(-E(\mathbf{a}|\mathbf{J})\right)$ is the partition function and ensures that all probabilities sum to one. For simplicity, here we label the set of fields and couplings as $\mathbf{J}$.

Given a sample of configurations, one may be interested in inferring back the model from which these samples were generated, or at least a model reproducing the statistical properties of such configurations, such as the one- and two-site frequencies, $f_i(a)$ and $f_{ij}(a, b)$. In general the Potts model defined above is the

simplest, or maximum entropy [35], probabilistic model capable of reproducing the observed frequencies. In the present case we know by construction that the Potts model is not only the simplest model to fit the data, but also the real model from which the sample was generated. To reproduce the statistics of the data, the parameters $h_i(a)$ and $J_{ij}(a,b)$ must be chosen such that site averages and correlations in the model match those in the data, i.e.,

$$\begin{aligned} \sum_{a_1,\dots a_N} \delta(a_i,a) P(a_1\dots,a_N|\mathbf{J}) &= f_i(a)\,, \\ \sum_{a_1,\dots a_N} \delta(a_i,a)\delta(a_j,b) P(a_1\dots,a_N|\mathbf{J}) &= f_{ij}(a,b)\,, \end{aligned} \tag{3}$$

where $\delta(a_i,a)$ is the Kronecker delta function, which is one if the symbol a_i at site i is equal to a and zero otherwise. Finding the parameters $h_i(a)$, $J_{ij}(a,b)$ that satisfy Eq. 3 constitutes the inverse Potts problem.

B. Cross-entropy and regularization

Formally, the solution to the inverse Potts problem is the set of fields and couplings that maximize the average log-likelihood or, equivalently, that minimize the cross-entropy between the data and the model. This cross-entropy can be written as

$$S(\mathbf{J}|\mathbf{f}) = \log Z(\mathbf{J}) - \sum_{i=1}^{N}\sum_{a=1}^{q} h_i(a) f_i^s(a) - \sum_{i=1}^{N-1}\sum_{j=i+1}^{N}\sum_{a=1}^{q}\sum_{b=1}^{q} J_{ij}(a,b) f_{ij}^s(a,b)\,, \tag{4}$$

where, for simplicity, we indicate the set of single and pairwise frequencies as $\mathbf{f}$ and the set of fields and couplings as $\mathbf{J}$.

To guarantee that the minimization of the cross-entropy is a well defined problem, a regularization term ΔS is added, which, n the Bayesian formulation, corresponds to prior knowledge over the parameters $\mathbf{J}$. A Gaussian prior distribution, also referred to as L_2-regularization, is a usual choice:

$$\Delta S = \gamma_h \sum_{i=1}^{N}\sum_{a=1}^{q} h_i(a)^2 + \gamma_J \sum_{i=1}^{N-1}\sum_{j=i+1}^{N}\sum_{a=1}^{q}\sum_{b=1}^{q} J_{ij}(a,b)^2. \tag{5}$$

The regularization parameters γ_J and γ_h are related to the prior variances of fields (σ_h^2) and couplings (σ_J^2) through $\gamma_h = 1/(B\sigma_h^2)$, and $\gamma_J = 1/(B\sigma_J^2)$, where B is the number of configurations in the sample. When the penalties are relatively weak ($\gamma \sim \mathcal{O}(1/B)$), this regularization can be thought of as a weakly informative prior [36] whose main purpose is to prevent pathologies in the inference.

C. Gauge invariance

The $N \cdot q$ frequencies $f_i(a)$ and $\frac{1}{2}N(N-1)q^2$ correlations $f_{ij}(a,b)$, with $i<j$, are related to each other: the former sum up to 1, while the latter have the frequencies as marginals. Therefore, not all constraints in Eq. 3 are independent and multiple sets of parameters give the same probability distribution. In the language of physics this over-parameterization of the model is referred to as *gauge invariance* and the choice of one particular parameter set among the equivalent ones as *gauge choice*. This gauge invariance reduces the number of free parameters in the Potts model to $q-1$ fields for each site and $(q-1)^2$ couplings for each pair of sites.

In particular, we can reparameterize the model without changing the probabilities by an arbitrary transformation of the form:

$$\begin{aligned} h_i(a) &\to h_i(a) + H_i + \sum_{j(\neq i)} K_{ij}(a) \\ J_{ij}(a,b) &\to J_{ij}(a,b) - K_{ij}(a) - K_{ji}(b) + \kappa_{ij} \end{aligned}$$

for any $K_{ij}(a)$, H_i and κ_{ij}. In the so called lattice-gas gauge, this freedom is used to define a *gauge state* c_i at each site such that

$$J_{ij}(a, c_j) = J_{ij}(c_i, b) = h_i(c_i) = 0\,, \tag{6}$$

for all states a, b and sites i, j. The couplings and fields are transformed as follows:

$$\begin{aligned} h_i(a) &\to h_i(a) - h_i(c_i) + \sum_{j\neq i} \left(J_{ij}(a, c_j) - J_{ij}(c_i, c_j)\right)\,, \\ J_{ij}(a, b) &\to J_{ij}(a, b) - J_{ij}(c_i, b) - J_{ij}(a, c_j) + J_{ij}(c_i, c_j)\,. \end{aligned} \tag{7}$$

Two common gauge states are the most and the least frequent states of each site, defining respectively the *consensus gauge* and the *least-frequent gauge*. In protein analysis, the gauge state is often fixed to the amino acid present at site i in a reference sequence, called *wild-type* sequence. An alternative choice is the so-called *zero-sum gauge*, in which

$$\sum_{c=1}^{q} J_{ij}(a, c) = \sum_{c=1}^{q} J_{ij}(c, a) = \sum_{c=1}^{q} h_i(c) = 0\,, \tag{8}$$

for all states a and all variables i, j. Fields and couplings can also be put in the so-called zero-sum gauge through

$$\begin{aligned} h_i(a) &\to h_i(a) - h_i(\cdot) + \sum_{j(\neq i)} \left[J_{ij}(a, \cdot) - J_{ij}(\cdot, \cdot)\right]\,, \\ J_{ij}(a, b) &\to J_{ij}(a, b) - J_{ij}(\cdot, b) - J_{ij}(a, \cdot) + J_{ij}(\cdot, \cdot)\,, \end{aligned} \tag{9}$$

where $g(\cdot)$ denotes the uniform average of $g(a)$ over all states a at fixed position.

Note that, while all observables such as the moments of the distribution are invariant with respect to the gauge choice, the fields and the couplings are not. Arbitrary functions of the couplings and fields, such as the commonly-used Frobenius norm of the couplings, are also not generally gauge invariant. If not explicitly stated, the comparisons shown in this paper are performed in the consensus gauge, but the choice of the gauge for the inference and for the analysis of the inferred network can be different. The gauge chosen during the inference will be further discussed in section II D in the descriptions of ACE and PLM.

D. Algorithms

The presence of the partition function Z in Eq. 4 precludes direct numerical minimization of the cross-entropy when the system size is large, since this requires summing over all $\prod_{i=1}^{N} q_i$ possible configurations of the system. However many approximate solutions have been proposed to tackle this issue. We briefly recall two of these methods to respectively approximate the cross-entropy or the log-likelihood: the Adaptive Cluster Expansion (ACE) and PseudoLikelihood Maximization (PLM).

1. *Adaptive cluster expansion (ACE)*

The cross-entropy (Eq. 4) can be exactly decomposed as a sum of cross-entropy contributions, calculated recursively (see Appendix VIII A and VIII B). The adaptive cluster expansion [2, 31, 32] is based on the idea of summing up cluster contributions based on their importance as quantified by their absolute contribution to the cross entropy. To this end an inclusion threshold parameter t is introduced and only clusters with cross-entropy contributions larger than the threshold t are included. The inclusion threshold t is then progressively decreased to include more and more clusters in the summation. The expansion is usually stopped when the frequencies and correlations of the inferred model reproduce the empirical ones to within the statistical error bars due to finite sampling. The inference routine which has been used in this paper is publicly available at `https://github.com/johnbarton/ACE` . For an input sample of size B, the regularization parameters are set to $\gamma_J = 1/B$ and $\gamma_h = 0.01/B$, corresponding to a variance of the prior distribution of couplings of order 1 and a variance of fields of order 100.

2. Pseudo-likelihood maximization (PLM)

The idea behind Pseudo-Likelihood Maximization is to approximate the full likelihood of the data given the model (or equivalently the full cross-entropy (Eq. 4) by the site-by-site maximization of the conditional probability of observing one state at a site, given the observed states on the other sites. This approximation makes the problem tractable, and it also makes possible to parallelize the computation for the different sites. Pseudolikelihood is a consistent estimator of the likelihood in the limit of infinite input data. For this study, a version of the asymmetric pseudolikelihood maximization [9, 28] capable of working with a site-dependent number of Potts states has been implemented adapting the public code by M. Ekenberg and E. Aurell at https://github.com/magnusekeberg/plmDCA.

Unlike ACE, the networks inferred by PLM with L_2-regularization are always fully connected. As empirically shown in protein sequence analysis [9, 22, 23] and in theoretical analyses [37], large regularization is needed in the presence of fully connected networks to avoid overfitting and thus to improve contact and fitness predictions. We have tested different regularization strengths, see Sec. III, and fixed $\gamma_J = N/B$, $\gamma_h = 0.1/B$ for system with N variables and input sampling of size B.

With PLM gauge invariance is automatically broken. The inference is performed in the gauge that minimizes the L_2-regularization:

$$\gamma_J \sum_{b=1}^{q_j} J_{ij}(a,b) = \gamma_h h_i(a) \ , \quad \gamma_J \sum_{a=1}^{q_i} J_{ij}(a,b) = \gamma_h h_j(b) \ , \quad \sum_{a=1}^{q_i} h_i(a) = 0 \ . \tag{10}$$

As for ACE, the PLM fields and couplings are subsequently transformed to the consensus gauge for comparison.

III. REGULARIZATIONS

A. Removing variable states

1. Color compression

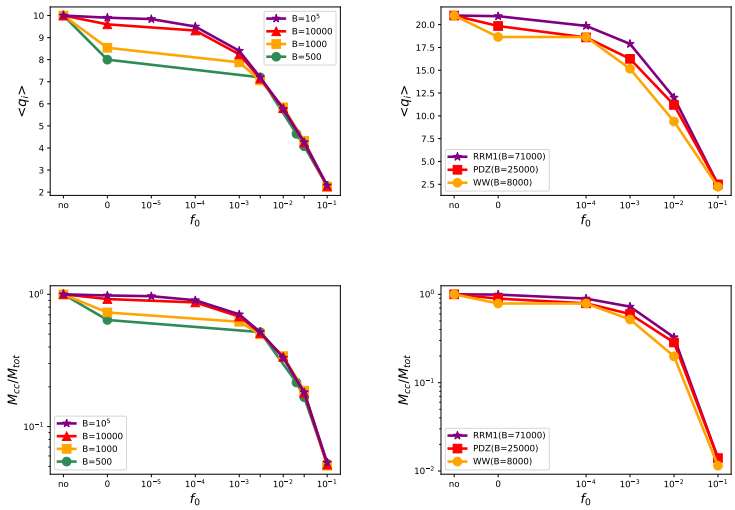
So far we have described (Eq. 4 and 5) how to infer the parameters of a Potts model where the number of states q is the same at all sites, but it is easy to generalize this procedure to Potts models in which the number of states q_i depends on the site i. This situation naturally arises due to sampling: states with very small probabilities are rarely observed. For instance, in multiple sequence alignments of protein families, for the large majority of sites. only a subset of the full $q = 21$ possible amino acids are observed. In the color compression procedure, for each site i, we model explicitly only the q_i states observed with a frequency $f_i(a)$ larger than a frequency cutoff threshold f_0

$$f_i(a) > f_0 \ , \tag{11}$$

and we group together the remaining $q - q_i$ low frequency states into a single one. The frequency of the grouped/compressed Potts state $q_i + 1$ is then the total frequency of the states that have been grouped together: $f_i(q_i + 1) \equiv \sum_{a'=q_i+1}^{q} f_i(a')$.

2. Artificial data sets on Erdős-Rényi random graphs

To benchmark color compressed inference we have generated synthetic data from a Potts model with $N = 50$ variables carrying $q = 10$ Potts states and interacting on an Erdős-Rényi random graph. To build the interaction graph, each edge in the network is included with probability 0.05, giving a mean connectivity 2.5, with a maximum connectivity equal to 7. An exemple of contact map is shown in Supplementary Fig. 16. The field and couplings parametes on interacting sites are selected from Gaussian distributions of mean $\mu = 0$ and standard deviations $\sigma_J^2 = 1$ and $\sigma_h^2 = 5$. Therefore if i and j interact J_{ij} is a 10×10 matrix whose elements are chosen independently according to the above distributions, and each element of the matrix is zero when the sites do not interact.



(a) Synthetic Data (b) Protein Sequence Data

FIG. 1: Mean number of colors per site $<q_i>$ (top) and relative reduction in the number of parameters to be inferred M_{cc}/M_{tot} (bottom), as a function of the f_0: no compression, 0 (only unseen symbols are removed from inference) and $1/B < f_0 < 0.1$. Data from a single ER realization (left), and sequence data for 3 protein families WW ($N = 31$), PDZ ($N = 84$), RRM ($N = 82$) (right) at different sample depths B.

We have generated 10 independent realizations of such ER models (networks of interactions and sets of fields and couplings). For each realization, $B = 5\ 10^2, 10^3, 10^4$, or 10^5 configurations are generated by Markov Chain Monte-Carlo sampling. The number of available data, $B \times N$, can be compared to the number of parameters to be inferred, $M_{tot} = qN + q^2N(N-1)/2 \simeq 1.2\ 10^5$. We have, for the previously listed values of B, $B \times N = 2.5\ 10^4, 5\ 10^4, 5\ 10^5$, or $5\ 10^6$, the first two being in heavy undersampling conditions, the third in scarce sampling, and only the last one being relatively well sampled.

3. *Reduction in the number of parameters in compressed data: comparison between ER data with $q = 10$ and sequence data with $q = 21$*

Tanks to color compression the number of colors per site, q_i, and thus the number of parameters $M_{cc} = \sum_i^N q_i + \sum_{i<j}^N q_i\, q_j$ are reduced with respect to the full number of colors per site, q, and the total number of parameters to be inferred M_{tot}. Fig. 1(left) illustrates how such reduction is achieved as a function of the compression threshold f_0 and for different sampling depths B for the artificial data. Increasing f_0 the inferred model goes from the full-color $\langle q_i \rangle = q$ Potts model to the Ising model with only two colors $\langle q_i \rangle = 2$ per sites indicating only the presence or absence of the most probable color on each site. The number of parameters to be inferred is divided by 50 at high compression.
Fig. 1 (right) shows a similar behavior for parameter reduction as a function of f_0 and sampling size B in protein sequence data, for three families that will be further described and analyzed in Sec. V. In the latter case the number of Potts states is the number of possible amino-acid types (plus the gap symbol), $q = 21$, and the number of parameters is decreased by 100 times at high compression. While protein sequence data are the direct application of color compression that we will consider in Sec. V, we have generated synthetic data from a Potts model with $q = 10$ instead of $q = 21$ states to speed up the numerical analysis performed varying sampling depth, compression threshold and inference methods; especially for the ACE algorithm, indeed, as it will be shown in Sec. VI, the computational cost of inferring a complete $q = 21$ Potts model may become very large. Moreover the aim of the present study is to investigate the interplay between sampling and effect of compression, and we do not expect qualitative changes between $q = 10$ and $q = 21$: we expect that the compression threshold f_0, at which the performances of the inferred model worsen with respect to the full Potts model does not depend on the maximal numbers of Potts state in the original model but rather on their occurrence in the sampling. At large compression threshold, we expect that model performance deteriorates

because highly frequent, well-sampled colors are grouped and their correlations are lost.

4. Color decompression

Once the restricted Potts model is inferred, we need to recover the complete model with q states at each site in order to compare it to the ground truth. To this aim, we have to determine the parameters for the states that were *grouped* on site i. We use the following procedure: For each grouped state a', the fields and the couplings are estimated through

$$h_i(a') = h_i(q_i+1) + \log\left(\frac{f_i(a')}{f_i(q_i+1)}\right) \ , \quad J_{ij}(a',b) = J_{ij}(q_i+1,b) \ , \tag{12}$$

where q_i+1 refers to the unique Potts state for the grouped symbols. This procedure allows us to correctly recover the local fields parameters reproducing the frequencies $f_i(a')$ for the grouped symbols, while a common coupling parameters is assigned to all the grouped symbols. Then, we associate fields and couplings to states that are never observed in the sampling, hereafter referred to as *unseen* states, through a natural extension of the procedure described above. We assign a pseudocount frequency $f_i(a") = \alpha/B$ to the unseen symbols ($a"$). Herafter, we have fixed $\alpha = 0.1$, in the expected range $0 < \alpha < 1$. When the grouped state is not present while unseens are present, we use the least probable state on the given site (l_i) instead of the grouped state in the denominator of Eq. 12, by replacing q_i+1 by l_i.

The choice of associating the unseen states to the grouped state or the least probable state is both simple and effective. Indeed, it follows the gauge choice, and yields fields with lower values than for the observed states. A detailed comparison of the effects of the pseudo-count above and of the standard L_2 regularization in the simple case of an independent model is presented in Appendix VIII E.

5. Gauge used in the ACE inference

ACE inference has been implemented in the lattice-gas gauge (Eq. 6) [32]. It is important to notice that the choice of the gauge symbol may have some effect on the performance of the inference procedure because the regularization term is not gauge invariant. For abundant data or in the limit of large compression, Potts states are well sampled and the choice of the gauge symbol is largely irrelevant. However, for few data or in the absence of color compression, or at small compression, the best performance is obtained by choosing as gauge symbol the least-probable Potts state on each site. In this way, all the fields and couplings corresponding to at least one poorly sampled state are fixed to zero, and have therefore null statistical variances by construction. Fields and couplings are then translated to the gauge in which the most probable symbol is chosen as gauge-symbol (consensus gauge) to perform the comparisons described in the next sections using Eq. 7. This consensus gauge is best for comparison because the statistics of the consensus symbols (on all sites) are the easiest ones to measure accurately.

B. Removing interactions

We also regularize the inference procedure by limiting the number of non-zero interactions. Sparsification of the interaction network is sometimes achieved through L_1 regularization of the couplings. Hereafter, we show that the inclusion threshold of the ACE inference procedure defined in Section II D 1 plays a similar role, while not affecting the amplitude of non-zero couplings.

1. Role of ACE inclusion threshold: sparse versus dense inferred graphs

Fig. 2 shows the behavior of the ACE algorithm as a function of the inclusion threshold t for one particular graph, hereafter called ER05, with $B = 1000$ sampled configurations, analyzed with a color compression of $f_0 = 0.01$. This representative data set will be our reference case. For each threshold t used to select clusters in the ACE expansion, the model frequencies $\langle\delta(a_i,a)\rangle$ and $\langle\delta(a_i,a)\delta(b_i,b)\rangle$ calculated by Monte-Carlo simulation are compared to the data frequencies $f_i(a)$ and $f_{ij}(a,b)$ (see Eq. 3).

As detailed in [2, 32], to monitor the ability of the inferred model's ability to reproduce the measured frequencies and correlations while avoiding overfitting, we define a relative error that is the ratio between the deviations of the predicted observables from the data, $\Delta f_i(a) = \langle \delta(a_i, a) \rangle - f_i(a)$ and $\Delta f_{ij}(a,b) = \langle \delta(a_i, a)\delta(b_i, b) \rangle - f_{ij}(a,b)$, and the expected statistical fluctuations due to finite sampling, $\sigma_i(a) = \sqrt{f_i(a)(1-f_i(a))/B}$ and $\sigma_{ij}(a,b) = \sqrt{f_{ij}(a,b)(1-f_{ij}(a,b))/B}$. The relative error on frequencies is

$$\epsilon_{\rm p} = \frac{1}{Nq}\sqrt{\sum_{i,a}\left(\frac{\Delta f_i(a)}{\sigma_i(a)}\right)^2}. \tag{13}$$

The relative error on connected correlations, $c_{ij}(a,b) = \langle \delta(a_i, a)\delta(b_i, b) \rangle - \langle \delta(a_i, a) \rangle \langle \delta(b_i, b) \rangle$, is

$$\epsilon_{\rm c} = \frac{2}{N(N-1)q^2}\sqrt{\sum_{i<j,a,b}\left(\frac{\Delta c_{i,j}(a,b)}{\sigma^c_{i,j}(a,b)}\right)^2}, \tag{14}$$

where we estimate the standard deviation in the connected correlations as $\sigma^c_{i,j}(a,b) = \sigma_{ij}(a,b) + f_j(b)\sigma_i(a) + f_i(a)\sigma_j(b)$. Finally, the maximum relative error is

$$\epsilon_{\rm max} = \max_{\{i,j,a,b\}} \frac{1}{\sqrt{2\log(M)}}\left(\frac{|\Delta f_i(a)|}{\sigma_i(a)}, \frac{|\Delta f_{ij}(a,b)|}{\sigma_{ij}(ab)}\right), \tag{15}$$

where $M = Nq + (N(N-1)/2)q^2$ is the total number of one- and two-point correlations. As shown in Fig. 2 (top panel) the relative errors defined above have a nonmonotonic behavior as a function of the threshold, reaching relative minima that successfully reconstruct the data ($\epsilon_{\rm max} < 5$) at multiple values (marked by asterisks) of the expansion threshold t, see Table I. The regularized cross entropy, the total number of clusters included in the expansion, and their maximal size as a function of the cluster inclusion threshold t are also shown Fig. 2.

The cluster inclusion threshold acts as an additional regularization. There are 3 plateaus in the regularized cross entropy as a function of the cluster inclusion threshold t of Fig. 2: the first plateau corresponds to an independent model, the second one to a sparse interaction network, and the third one to a fully connected network. The number of edges present in the inferred graph of Fig. 2 is given by the number N_2 of 2-site clusters in the ACE expansion and is shown in Table I for the threshold corresponding to the minimal relative errors ϵ_{max}. In particular the two relative minima better reproducing the data correspond to 2 different inferred networks. The minimum with $\epsilon_{\rm max} = 3.9$ is at high threshold ($t = 0.036$) and is characterized by a numbers of edges $N_2 = 55$ smaller than the total number $N(N-1)/2 = 1225$ of possible pairs. The inferred model is therefore a sparse graph, with a number of edges N_2 comparable with the number of edges of the model used to generate the data ($N_0 = 59$ for the model used to generate the data in Fig. 2). The second relative minimum with $\epsilon_{\rm max} = 1$ is at low threshold $t = 6.4 \times 10^{-5}$ where the expansion includes the maximal number of 2 site clusters $N_2 = N \times (N-1)/2$, corresponding to a fully connected graph. As can be guessed by the difference in the connectivity between the original and inferred model, and as we will better quantify in Section IV A, the fully connected solution is overfitting the data.

2. *spACE, a variant of ACE for sparse interaction networks inference*

To force the ACE algorithm towards a sparse solution we introduced a new procedure in the cluster expansion (available at https://github.com/johnbarton/ACE)„ which stops the algorithm at a maximal number N_2^{max} of 2-site clusters, see Appendix VIII B. This procedures imposes a prior knowledge on the sparsity of the interaction graph by giving an upper bound for the number of edges. The Erdős-Rényi random graph models used here to generate the data have an average connectivity of 2.5 neighbors per site, so we can use this prior knowledge to fix a maximal number of edges to $N_2^{max} = N \times 2.5 \approx 125$. In practice to find the best sparse graph with a number of edges smaller than the prescribed value N_2^{max} we spanned the threshold range in regular intervals [38], at which the local minima of the absolute error and the corresponding t parameters are recorded (see Table I) The parameters corresponding to the global minimum (under the sparsity conditions) are then chosen ($t = 3.6 \times 10^{-2}$ for Table I). The spACE procedure is stable when changing N_2^{max} around its prior fixed average value. In the following we have stopped the algorithm

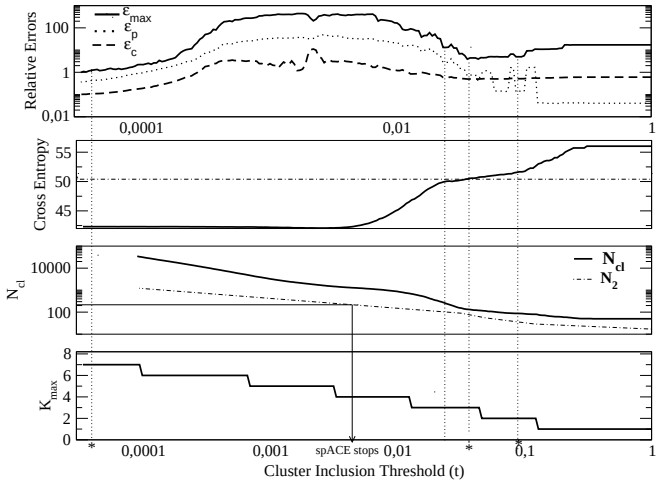


FIG. 2: Cluster expansion as function of the cluster inclusion threshold t for one data set, obtained by sampling one realization of an Erdős-Rényi random graph with $B = 1000$ configurations and applying ACE to compressed data with $f_0 = 0.01$. From top to bottom: *i)* relative reconstruction errors versus t on frequencies ϵ_p, connected correlations ϵ_c and maximal relative error ϵ_{max}. Stars indicate possible solutions of the inverse models, reproducing well the empirical 1- and 2-point correlations ($\epsilon_{\max} < 5$), which are obtained at different expansion thresholds and corresponding to sparse graphs at large thresholds, or fully connected reconstructed graphs at small thresholds, see Table I *ii)* Regularized cross entropy vs. t, *iii)* number of total clusters N_{cl} and number of 2-site clusters N_2 included in the expansion vs. t, *iv)* maximal cluster size versus t. The spACE expansion stops at $N_2 = 200$.

t	ϵ_{max}	N_2	S_c^λ	S_c
1	17	0	56	56
9.6×10^{-2}	4.9*	29	51.7	51.2
3.6×10^{-2}	3.9*	55	50.6	49.9
2.2×10^{-2}	34	90	49.7	48.8
3.4×10^{-5}	1*	1225	42.3	39

TABLE I: Inferred models obtained varying the expansion thresholds for the reference data (ER05,B=1000 configurations), of Fig. 2 , with compression frequency $f_0 = 0.01$. The Table gives the cluster inclusion threshold t, the number N_2 of 2-site clusters, the maximal relative error ϵ_{max}, the regularized cross entropy S_c^λ and the cross entropy S_c obtained with the cluster expansion at different thresholds t. The entropy of the model having generated the data is $S = 50.3$. Possible solutions ($\epsilon_{\max} < 5$) are indicated by asterisks. The optimal threshold determined by the spACE procedure is 3.6×10^{-2}.

for $N_2^{max} = 200$ (shown on Fig. 2), after checking that the value $N_2^{max} = 100$ gave similar results. This procedure greatly reduces the computational time, which increases linearly with the number of computed clusters and grows exponentially, as $q^{K_{max}}$, with their maximal size K_{max} (see Sec. VI and Appendix VIII C): as shown in Fig. 2 $K_{max} = 3$ at $N_2^{max} = 200$ while $K_{max} = 7$ for the fully connected graph. The number of inferred parameters M_{cc} in Fig. 1 is further reduced by a factor 5 10^{-2} for the best sparse model on ER data in Table I. having couplings only on connected sites.

IV. BENCHMARKING OF COLOR COMPRESSION AND DECOMPRESSION ON SYNTHETIC DATA

To carry out an extensive analysis of the effects of the color compression introduced in Section III on the quality of the inference, we will apply it to artificial data generated by Potts models on Erdős-Rényi (ER) random graphs. The model and the generation of the data are described in Section III A 2.

Once the data are obtained, we apply the compression schemes introduced in Section III with no color compression and with frequency cut-off $f_0 = [0, 10^{-4}, 10^{-3}, 3 \cdot 10^{-3}, 10^{-2}, 3 \cdot 10^{-2}, 10^{-1}]$. Note that all frequency thresholds $0 < f_0 \leqslant 1/B$ give the same color compression, so we infer the model only for the upper value in this range and thus the number of the tested frequency thresholds depends on B. Moreover, $f_0 = 0$ corresponds to removing from the inference only the unseen states.

Given the 10 realizations of the Erdős-Rényi model, the 4 sample sizes and the 5 to 8 (depending on the sampling) values of the frequency threshold define 280 data sets. For each of them, we have inferred the corresponding Potts parameters, both with the ACE and the PLM algorithms.

A. Probability distributions

The Kullback-Leibler (KL) divergence measures how the inferred probability distribution of the possible configurations diverges from the empirical one (defined from the data samples), and can be computed as:

$$D(P_{\boldsymbol{J}^{real}}||P_{\boldsymbol{J}^{inf}}) = \sum_{\mathbf{a}} P_{\boldsymbol{J}^{real}}(\mathbf{a}) \log \frac{P_{\boldsymbol{J}^{real}}(\mathbf{a})}{P_{\boldsymbol{J}^{inf}}(\mathbf{a})} = \log(Z_{inf}) - \log(Z_{real}) + \langle E_{inf}(\mathbf{a}) - E_{real}(\mathbf{a}) \rangle_{\mathbf{a}\, in\, real} \ ,$$

where $\mathbf{a} = \{a_1, \dots a_N\}$ is a configuration and $\langle \cdot \rangle_{\mathbf{a}\, in\, real}$ indicates the average over the configurations generated by Markov Chain Monte Carlo (MCMC) from the real model. The first and the second lines are identical only when an infinite configuration sample is employed. Here, we estimate the average over $P_{\boldsymbol{J}^{real}}$ using an ensemble of 50,000 MCMC configurations sampled from the model.

As described before, the computation of the partition function Z is far from being trivial, and was done in two ways. First, we used Annealed Importance Sampling (AIS) [39], starting from the independent-site model: All initial couplings were set to zero, while initial fields were computed as $h_i^0(a) = \log(f_i(a) + \alpha) - \log(f_i(cons_i))$ where $cons_i$ is the most common state at site i and $\alpha = 1/B$ is the smallest observed frequency used as regularization. Then a chain of models with increasing couplings (up to the inferred values) are thermalized and the ratios of their partition functions may be estimated. Secondly, the Kullback-Leibler divergence and the logarithm of the partition function can also be directly estimated by the ACE procedure (Table I), see Appendix VIII A. The KL divergences obtained directly from the ACE expansion and the ones obtained with importance sampling are very similar, as shown in Table II, for the reference case in Fig. 2 at the optimal cluster inclusion threshold corresponding to a sparse inferred network. The values of the logarithm of the partition function, and of the entropy are also consistent between the two methods. In the following we will use the annealed importance sampling to calculate the KL divergence to compare results from PLM and ACE.

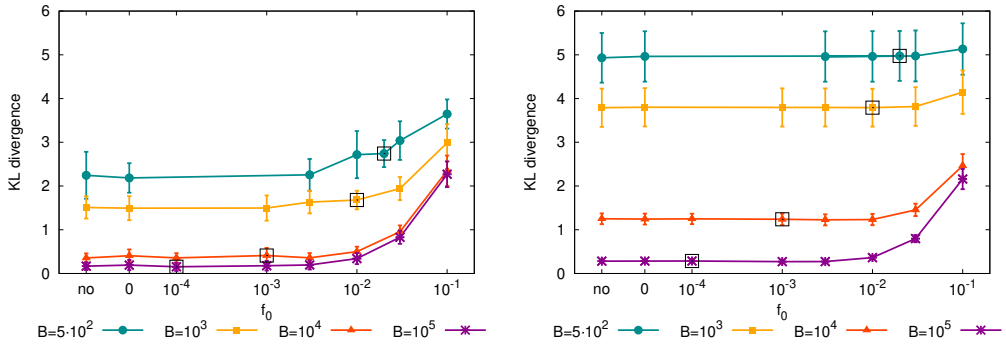
1. *KL Divergence for ACE models at different inclusion thresholds t*

Table II displays the KL divergences for the reference data set and different cluster inclusion thresholds of Fig. 2 and in table I obtained both with importance sampling and the ACE expansion.

The fully-connected-graph model has a larger KL divergence and is therefore overfitting the data, while the sparse-graph model reproduces better the original model. All results for the ACE expansion presented in the following are obtained with the spACE procedure to infer a sparse graph. For the fully connected solution, due to overfitting, the cross entropy of Table I is not a good approximation to the entropy. Therefore, the estimate of the logarithm of the partition function and of the entropy given in table II are significantly different from the ones obtained by the AIS method.

cluster inclusion threshold t	logZ(AIS)	logZ(ACE)	S(AIS)	S(ACE)	KL(AIS)	KL(ACE)
1	28.9	28.6	56.9	56	6.4	6
9.6×10^{-2}	32	31.9	52.6	51.4	2.2	2.3
3.6×10^{-2}	32.5	31.8	51.8	50.7	1.5	1.5
2.2×10^{-2}	34.2	32.8	52.9	50.9	2.5	3
3.4×10^{-5}	36	26	57.8	49.1	7.4	5

TABLE II: Comparison between importance sampling (AIS) and ACE methods to obtain the logarithm of the partition function Z, the entropy S, and the Kullback-Leibler divergence (KL), at the different sparsity threshold t for the reference model, ER05, data sampling: B=1000 and color compression $f_o = 0.01$, the optimal threshold determined by the spACE procedure is 3.6×10^{-2}. Fluctuations of the above values in AIS estimation over repetition of MC sampling are of the order $5\,10^{-3}$ for the sparse models and $5\,10^{-2}$ for the fully connected model.



(a) ACE (b) PLM

FIG. 3: Kullback-Leibler divergence between real and inferred probability distributions averaged over 10 realizations plotted as function of the compression parameter f_0 for different sample sizes. The left plot is for ACE, the right plot for PLM. Error bars are standard deviations over the 10 realizations. Black empty squares correspond to the reference cutoff frequency $f_0^* = 10/B$.

2. *KL Divergence as a function of the sampling depth B and the color compression threshold f_0*

Fig. 3 shows the mean over the ten ER realizations of the KL divergence between the real and the inferred distributions for various sampling depths and compression parameters for both ACE (left) and PLM (right). As expected, the KL divergence decreases for bigger samples, becoming very close to zero for $B = 10^5$. The same happens for the standard deviations over the 10 realizations. ACE gives smaller KL divergences with respect to PLM, showing that the sparsity imposed in the spACE procedure gives a model reproducing better the original, sparse ER models.

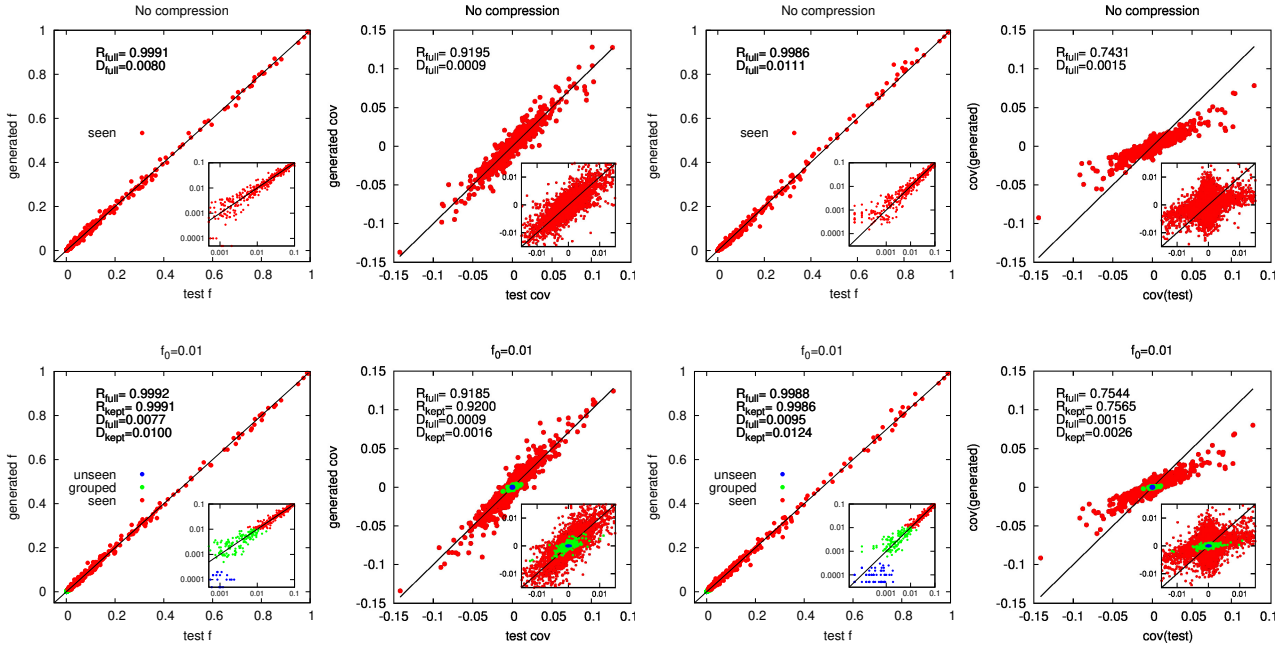
The black empty squares in Fig. 3 indicate a reference compression threshold $f_0^* = 10/B$, *i.e.* grouping symbols observed less than 10 times. We consider that frequencies larger than f_0 are reliably estimated [40] For $f_0 > f_0^*$ the compression procedure discards information on 'reliable' colors, and a loss in performance is expected. For both inference procedures the increase in KL divergence in Fig. 3 becomes visible only at a large value of the cut frequency $f_0 \gtrsim 0.001$, independently from the sampling depth. We indeed expect that the increase of the KL divergence due to the color compression procedure be a monotonic function of the cut frequency f_0, so it is irrilevant at small cut frequency. For high values of f_0 the increase of the KL divergence with respect to the uncompressed model is of course much more significant at high B.

Table II shows that the strong regularization ($\gamma_J \approx N/B$) used in the fully connected PLM inference is essential to reduce overfitting: a PLM model inferred with a weak regularization ($\gamma_J \approx 1/B$) gives indeed very large KL divergences, especially at low sampling depth. Color compression, acting as an additional

B	PLM ($\gamma_J = \frac{1}{B}, \gamma_h = \frac{0.002}{B}$)	PLM ($\gamma_J = \frac{50}{B}, \gamma_h = \frac{0.1}{B}$)	ACE ($\gamma_J = \frac{1}{B}, \gamma_h = \frac{0.01}{B}$)
10^3	12.97	3.80	1.35
10^4	2.85	1.28	0.37
10^5	2.10	0.30	0.18

TABLE III: KL divergences between inferred and empirical distributions on the reference data sets at different B and with no color compression. The value obtained by weakly regularized PLM $\gamma_J = \frac{1}{B}$ is compared to the ones obtained by the standards strongly regularized PLM inference and weakly regularized spACE inference, used in Fig. 3.

regularization, helps in this latter case to reduce overfitting and to lower the KL divergence. As shown in Appendix VIII E 1 and Fig.15 the KL divergences for weak regularized PLM model reach indeed a minimal value, at intermediates or large compression thresholds f_0 depending on sampling depth, compatible, but slightly larger, than the one obtained for the usual strongly regularized PLM of Table II.



(a) ACE (b) PLM

FIG. 4: Reconstruction of average frequencies and covariances for the reference case (ER05, B=1000). Comparison between generated data and test set for no color compression (top) and $f_0 = 0.01$ for ACE (left,a) and PLM (right, b). The Pearson correlation coefficient (R) and the absolute error (Δ, Eq. 16) are marked on top of the plots both for the full model and the for the reduced one (only explicitly modeled states).

B. Low-order Statistics

In this Section we discuss the generative properties of the inferred models, in particular its ability to reproduce the low order statistics of the original model : the site frequencies $f_i(a) = \sum_{generated\ \mathbf{a}} \delta(a_i, a)/B_{gen}$

and covariances $cov_{ij}(a,b) = \sum_{generated\ \mathbf{a}} [\delta(a_i,a)\delta(a_j,b)/B_{gen}] - f_i(a)f_j(b)$. To benchmark the generative power of the inferred model as a function of the color compression two sets of 20000 configurations are generated by Markov Chain Monte Carlo, respectively with the real and with the inferred model for each B, f_0, and graph realization, and their low order statistics are compared.

Figure 4 shows, for the models inferred from the reference data set, the comparison of the frequencies and covariances computed from the configurations generated by the real model (test sequences) and by the models inferred with ACE and PLM, without color compression (top panels) and with $f_0 = 0.01$ (bottom panels). As shown in the figures PLM covariances are downscaled, due to the strong regularization, as happens for the couplings (Section IV D). Moreover PLM assigns smaller frequencies to the unseen Potts states (left panels of Fig. 4 in log-log scale). This is probably due to the fact that the pseudocount used during decompression seems to be well fixed for the weak regularization used in ACE but not for the large regularization used in PLM. The insets in Fig 4 show that, contrary to spACE, zero covariances are set to non-zero values with PLM due to overfitting.

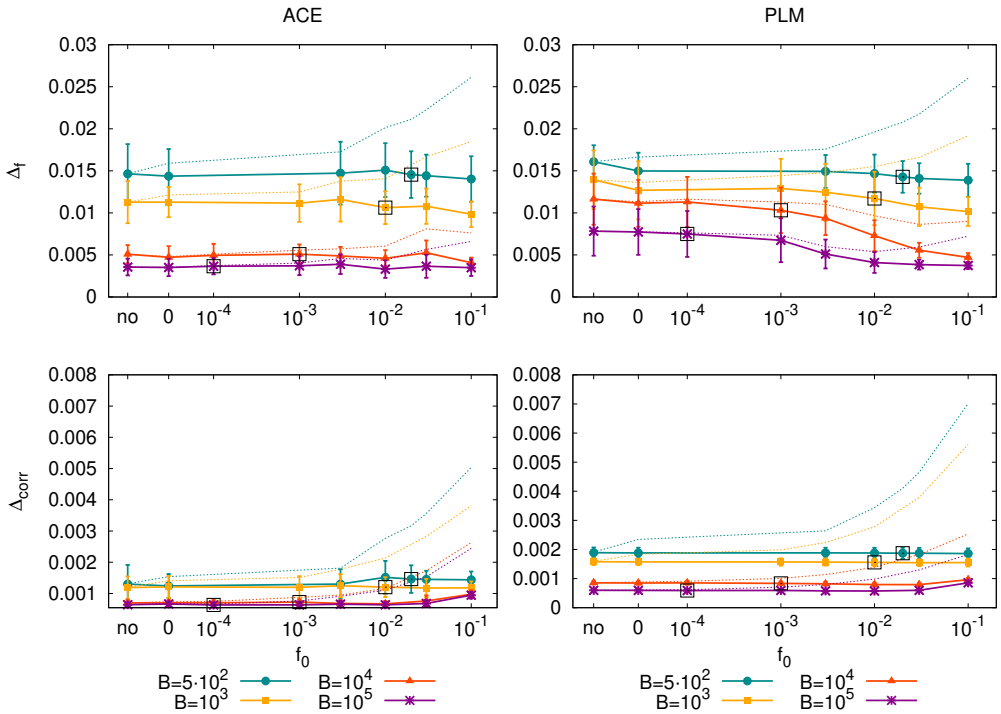


FIG. 5: Absolute error (Eq. 16) of frequencies (Δf, top panels) and covariances (Δcorr, bottom panels) averaged over 10 ER realizations, as a function of the color compression for several sample sizes. Dashed lines: error on explicitly modeled Potts states only. Full lines: error on all parameters. Error-bars are standard deviations computed over the 10 realizations. Inference is performed respectively by ACE (left) and PLM (right).

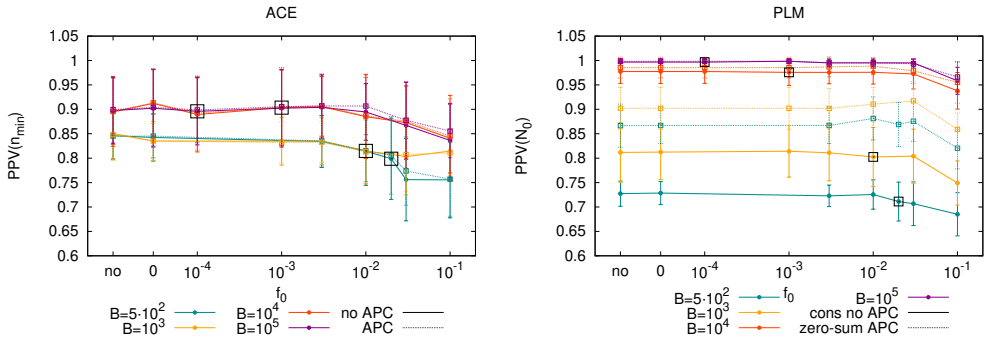
To have a more systematic comparison, we analyzed their absolute mean square error defined as:

$$\Delta f = \sqrt{\frac{\sum_i \sum_a \left(f_i^{gen}(a) - f_i^{test}(a)\right)^2}{\sum_i q_i}}\ , \quad \Delta\text{corr} = \sqrt{\frac{\sum_{ij}\sum_{ab}(cov_{ij}^{test}(a,b) - cov_{ij}^{gen}(a,b))^2}{\sum_{ij} q_i \cdot q_j}}\ . \tag{16}$$

These quantities are then computed for different B and f_0 and averaged over 10 graph realizations.

Figure 5 shows the Mean Square Errors (MSE) for the frequencies and covariances. spACE has again better performances than PLM [41]. The MSE for the full Potts model (full lines) shows a little dependence on the compression frequency f_0; meanwhile the MSE restricted to explicitly modeled colors (dotted lines) generally increases at large compression performances. The only exception is that frequencies are better

reproduced with PLM at strong color compression because the decompression procedure (Eq. 12), acting as an independent model, correctly assigns fields to grouped states. The above observations are simply explained by the fact that the contribution to the MSE of a color increases with its frequency. The MSEs on the full Potts model, after decompression (full lines in Fig. 5) are quite independent from f_0, as the averages in the MSE are dominated by the large amount of colors with small frequencies, see Fig. 4. On the contrary the MSE on explicitly modeled colors (dot lines in Fig. 5) shows a large increase with the compression threshold because the mean is restricted to more and more frequent colors.



(a) ACE's $PPV(n_{min})$ (b) PLM $PPV(N_0)$

FIG. 6: ACE and PLM interaction graph reconstruction as a function of the color compression f_0. Positive Predicted value $PPV(n = min\{N_0, N_{pred}\})$ for ACE and PLM inference. Points and error bars are averages and standard deviations obtained on 10 ER realizations. For ACE all quantities are computed in the consensus gauge with (dotted line) and without (full line) APC. (Left and middle plot). Right Plot $PPV(N_0)$ with PLM either in the consensus gauge without APC (full line), or in the zero-sum gauge with APC (dotted line).

C. Interaction networks

In this section we focus on the reconstruction of the interaction network and the prediction of pairs of sites that are connected, or "in contact", in the interaction graph. The original ER graph is sparse, with an average connectivity of about 2.5. We can predict contacting sites as those site pairs with large couplings, as traditionally done for protein structures [13–15]. To this end, we compute the Frobenius norm of the (10×10) inferred and decompressed coupling matrix between each pair of sites i, j, $F_{ij} = \sqrt{\sum_{a,b} J_{ij}(a,b)^2}$.

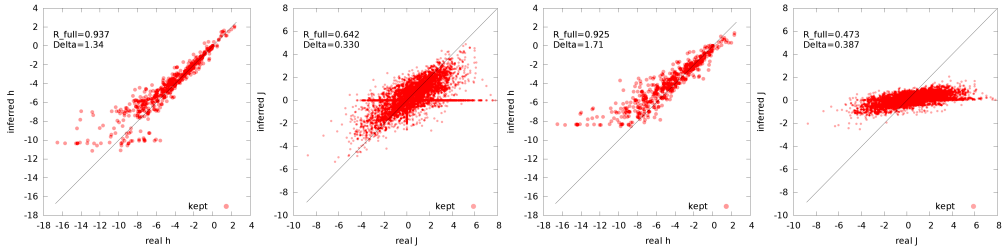
The heat map of the Frobenius norms of the couplings inferred by ACE and PLM at different color compression gives very similar results and allow us to identify the largely coupled sites by ACE and PLM (see Figs. 16 in Appendix). The only difference between ACE and PLM is that spACE infer a sparse network with zero Frobenius norms on the majority of links, while PLM with the L_2 norm regularization described in Eq. 5 infers a dense Frobenius norm matrix, There is therefore no straightforward separation between pairs of sites predicted to be in interaction or not.

To gain more insight into these predictions, as done for protein structure [23, 28, 42], we can sort links by decreasing Frobenius norm and follow the precision obtained progressively including the corresponding links in the so called Positive Predicted Value (PPV) curve (see Appendix VIII G).This is shown in Figs. 6 for a number of links up to the last inferred one for ACE or to $N_0 = 50$ for PLM, for the reference case.

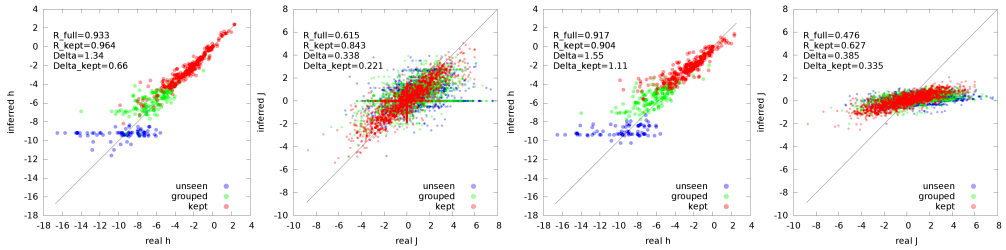
The Frobenius norm is computed in the consensus gauge, which will be used to compare the couplings and fields in the next section and in the zero sum gauge used on protein structure prediction [18, 23, 43]. Performance can be further improved with the Average Product Correction (APC) (defined in Appendix VIII G) when using zero sum gauge. When inferring sparse networks, APC only corrects the ranking of the predictions in the PPV curve, but it does not change the overall number of site pairs predicted to be in interaction, nor the global precision, on the contrary APC on the zero sum gauge largely improve PLM precision, as expected. Figure 6 (bottom) shows the average PPV, over all the data realization and as a

function of f_0, at the number of real links $PPV(N_0 = 50)$ for PLM and at the lesser between the predicted N_{pred} or real N_0 number of links with ACE.

As expected, the plots show that the contact prediction improves with sampling, and that the APC significantly improves the results for PLM in zero sum gauge. PLM generally gives higher PPV, especially at high sampling depth $B = 10^4$ and $B = 10^5$ where the reconstruction error for spACE are due to the sparsity threshold. We have verified that, at large sample and when inferring fully connected netwoeks ACE has the same PPV as PLM. We see that, for both algorithms, the performance is usually stable against the introduction of color compression.



(a) ACE: no color compression (b) PLM: no color compression



(c) ACE: $f_0 = 0.01 = 10/B$ (d) PLM: $f_0 = 0.01 = 10/B$

FIG. 7: Comparison of inferred and real fields and couplings with ACE and PLM for one realization of ER graph with no color compression (top) and $f_0 = 0.01$ (bottom), for B=1000 sampled configurations. Parameters on explicitly modeled (kept), grouped, and unseen Potts states are colored differently. Left: field comparison. Right: coupling comparison. On top of each plot, the Pearson correlation coefficient (R) and the absolute error (Δ, as in Eq. 17) are indicated.

D. Couplings and Fields

In Fig. 7 we compare the fields and the couplings of the real model (x-axis) and the inferred Potts model (y-axis) obtained for the reference data of the graph ER05 sampled at $B = 1000$ without color compression (top panels) and with $f_0 = 0.01 = 10/B$ (bottom panels), respectively, with ACE and PLM. Different colors in Fig. 7 show Potts states (or Potts states pairs) occurring at different frequencies and therefore treated in the color compression procedure as explicitly modeled, grouped, or unseen in the configuration sample. For couplings, if at least one of the two Potts state is unseen, the pair is considered as unseen; if at least one site is grouped, the pair is considered as grouped; if both sites are explicitly modeled, the pair is considered as explicitly modeled. The comparison is performed in the consensus gauge.

Figure 7 shows that, as observed in Section IV C the sparse procedure, spACE, misses some edges and the corresponding couplings are fixed to zero. PLM couplings are systematically smaller in amplitude than real ones, ending up in a tilted entry-by-entry comparison. This is due to the large regularization introduced to avoid overfitting.

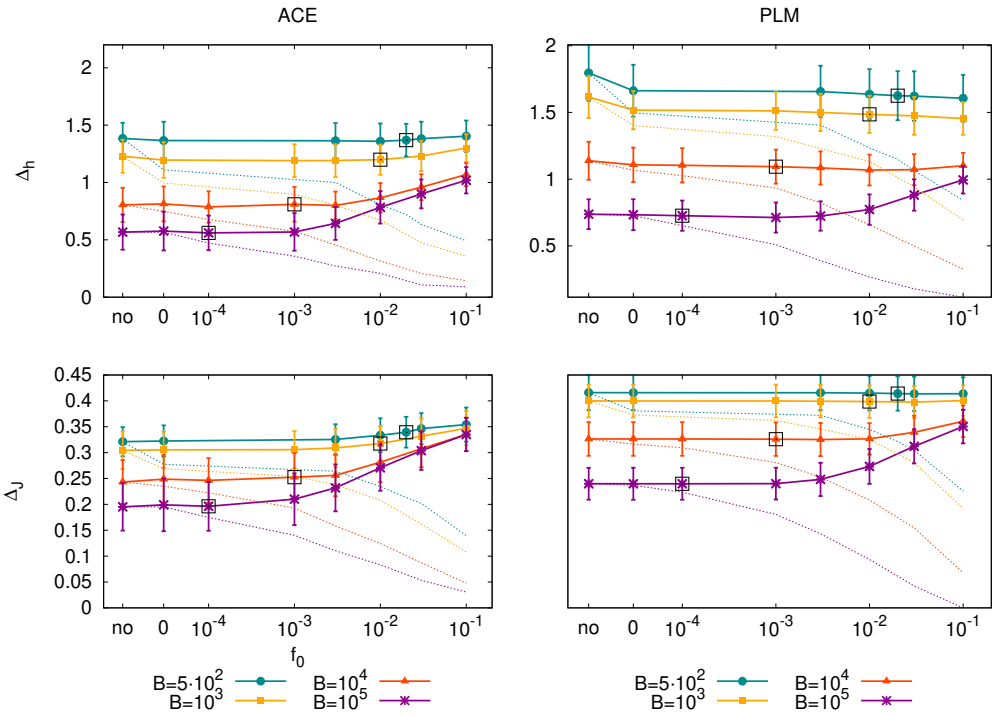


FIG. 8: Absolute errors (Eq. 17) on fields (Δh, top panels) and couplings (ΔJ, bottom panels) averaged over 10 ER realizations, as a function of the color compression for several sample sizes. Dashed lines: error on parameters related to explicitly modeled Potts states. Full lines: error on all parameters, after decompression of unseen and grouped Potts states (see Sec. III). Error-bars are standard deviations computed over the 10 realizations. Inference is performed respectively by ACE (left) and PLM (right).

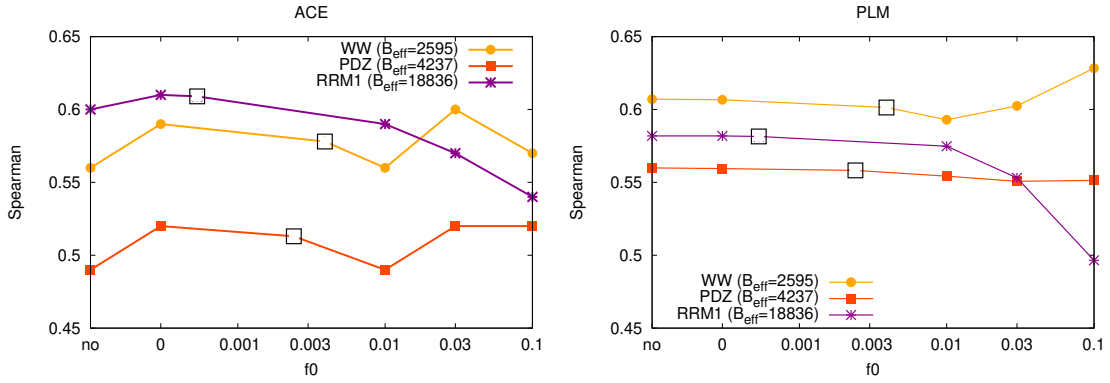
Figure 8 shows the absolute errors on the fields and couplings defined through

$$\Delta h = \sqrt{\frac{\sum_i \sum_a \left(h_i^{inf}(a) - h_i^{real}(a)\right)^2}{\sum_i q_i}} \;, \quad \Delta J = \sqrt{\frac{\sum_{ij} \sum_{ab} (J_{ij}^{real}(a,b) - J_{ij}^{inf}(a,b))^2}{\sum_{ij} q_i \cdot q_j}} \;; \tag{17}$$

These errors measure the average distances from the diagonal of the points in the scatter plots of Fig. 7. We observe that the couplings and fields reconstruction performances are stable as a function of the color compression up to the reference value $f_0^* \simeq \frac{10}{B}$, where they drop because the compression become too strong. Such drop is perfectly clear for the couplings parameters, especially within the small regularization used in spACE. Large compression threshold, as well as large coupling regularizations, degrades indeed the informations on the correlations of grouped, badly sampled, variables. The increase of field reconstruction error at all compressions is negligible for small sampling depth $B = 500, B = 1000$ and for both algorithm, showing that the decompression procedure introduced in Sec. III, correctly assign, in the limit of the available information, the large and negative fields for the grouped and unseen Potts symbols, as done by using prior information with the L_2 regularization and shown in Fig. 7. The dashed lines in Fig. 8, in agreement with Fig. 7, show that by restricting the coupling comparison to the explicitly modeled symbols, the better and better sampled ones at larger and larger f_0, the reconstruction indicators are better and better. In other words, even in the largely undersampled regime, parameters for well sampled colors are correctly inferred and are not affected by the poorly sampled states. This underlines the difference between sites and states in a Potts model. In the standard renormalization procedure [44] when the number of sites are reduced in an effective "renormalized" Potts model the parameter values of the retained sites change. In contrast, in the space of Potts states, grouping some of them, and keeping the probabilities conserved, does not affect the others [45].

V. APPLICATION TO SEQUENCE-BASED MODELS OF PROTEIN FAMILIES

We now apply our inference approach to protein sequence data. Input samples are multiple sequence alignments of protein families, the nodes of the graph are the protein sites, and states are the 20 amino acids plus the insertion-deletion symbol ($q = 21$). In this context, we aim at reconstructing the contact map [14, 42] or the fitness landscape [19–22]. In particular, we would like to compare the change of energy corresponding to single point mutations with respect to a wild-type protein sequence to the experimentally measured changes of fitness of the protein. Differently from ER data, sequences are not uniformly sampled. A reweighting procedure, described in Appendix VIII H, is usually introduced to reduce the initial number B of sequences in the sample to an effective number of independent ones B_{eff}.



(a) ACE (b) PLM

FIG. 9: Spearman correlation coefficient between the fitness predictions and the experimental measures found in literature, as a function of the color compression for ACE and PLM. The protein families used here are: WW (PF00397, in orange), PDZ (PF00595, in green) and RRM1 (PF00076 in light blue)

We here consider three protein families, whose fitnesses have been systematically assessed against single-point mutations. We specify for each of them the number of sites in the alignment N, the number of sequences B and the number of effective sequences B_{eff}:

- WW ($N = 31$, $B = 8251$, $B_{eff} = 2590$) is a protein domain that mediates specific interactions with protein ligands. Here fitness has been measured in terms of the capability to bind a certain ligand [46].
- PDZ ($N = 84$, $B = 24795$, $B_{eff} = 4240$) is a protein domain present in signaling proteins. Here fitness has been measured in terms of binding affinity [47];
- RRM ($N = 82$, $B = 70780$, $B_{eff} = 18800$)is an RNA recognition motif; fitness was estimated through growth rate measurements in [48].

Alignments and experimental fitness measures used in this section were taken from [22]. The PLM procedure was applied with the same large regularization used for the artificial data, $\gamma_J = N/B$, $\gamma_h = 0.1/B$. The SpACE procedure was applied with a threshold cutoff $N_2^{max} = 3N$, and a smaller regularization $\gamma_J = 10/B$, $\gamma_h = 0.1/B$; however the relative error ϵ_{max} (Eq.15) was generally too large even in its local minima, indicating that the procedure had not converged. As shown in [32] a Boltzmann Machine Learning (BML) procedure was further used, starting from the spACE inferred parameters as an initial guess, to better reproduce the low order statistics of the data and therefore the quality of the inferred model. Mutational cost of single mutations is predicted by their energetic cost, corresponding to the value of the field of the mutated amino acid, after having transformed the field parameters in the gauge of the wild type sequence through Eq. 7 [23].

Contrary to what happens for synthetic data, where the true model is known, the relationship between inferred energies and experimental fitness values may be nonlinear, so we use as a quality measure of the inference the Spearman correlation coefficient between them rather than the Pearson. The Spearman values of Fig. 9 for the full model (no compression) and PLM are in agreement to the ones previously obtained

[22] with PLM algorithm and a slightly different procedure, giving Spearman values of 0.6, 0.57, 0.5 for RRM1, WW and PDZ respectively. ACE + BML performances are compatible with PLM results, better for families with a large number of sequences such as RRM, and slightly worse in the PDZ case. The impact of color compression on the prediction is then investigated and shown in Fig. 9 and has to be compared to the quality of field reconstruction in Fig. 8, after decompression (full lines). For the WW and PDZ families, as for PLM with the ER data with $B = 1000$ and $B = 500$, the predictions are not affected by large compression thresholds. Very large compressions may even improve performances, as for the to WW family, where inference with almost binary Potts variables (at $f_0 = 0.1 < q_i >= 2.5$ see Fig. 1) give better predictions than with a full $q = 21$-state Potts model. A similar effect was observed also for HIV fitness predictions in [20]. The fact that large color-compressions can improve fitness predictions may be related to the experimental uncertainties and limited resolution on the fitness measures and also to the difficulty in estimating the effective number of independent sequences in the data. For the better sampled *RRM* family, predictions worsen at large compression $f_0 > 10/B_{eff}$, as expected and observed in Fig. 8 for the well sampled artificial data (at $B = 10^4$ and $B = 10^5$).

VI. GAIN IN COMPUTATIONAL TIME FOR SYNTHETIC AND PROTEIN SEQUENCE DATA

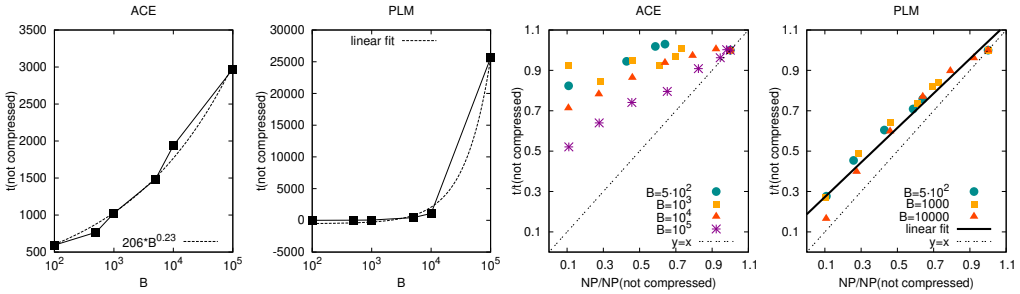


FIG. 10: Computational times averaged on all ER data for ACE and PLM for different B and color compression. Left: computational time, in seconds, using only 1 CPU, for ACE and PLM as a function B; a power law/ linear fits are added to ACE /PLM (dashed line). Right: computational time ratio between the compressed and decompressed inference for ACE (1 CPU) and PLM (25 parallel CPU) as a function of fraction of parameters to infer; dotted line: x=y; full line: linear PLM fit.

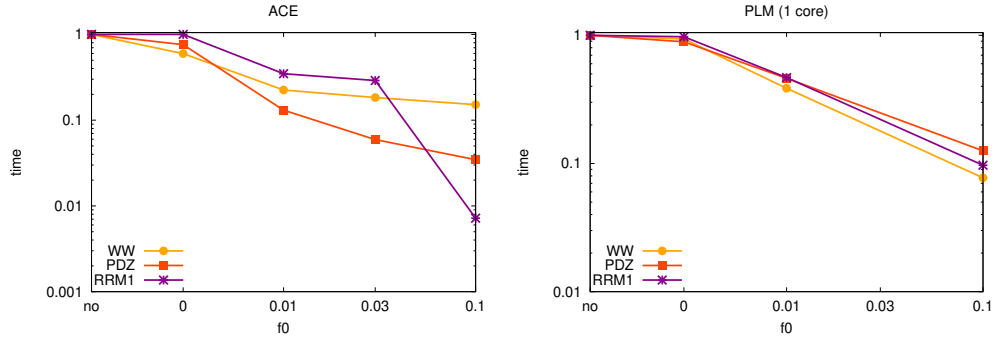


FIG. 11: Time gain due to color compression on protein sequences data with ACE (Left) and PLM (Right). The protein families used here are: WW (PF00397), PDZ (PF00595) and RRM1 (PF00076).

In this section we study how the computational time scales with the sample size and the color compression frequency threshold for spACE and PLM on synthetic and protein sequence data. Times have been obtained on a processor Intel® Xeon(R) CPU E5-2690 v4 @ 2.60GHz x 56. The limiting factors for computational time with the ACE expansion are three: the number of colors $< q >$ in the inferred model, due to the

	$B=10^3$						$B=10^5$					
$f0$	t_{ACE}	$\hat{N}_{cl}$	k_{max}	$< q_i >$	t^1_{PLM}	t^{20}_{PLM}	t_{ACE}	$\hat{N}_{cl}$	k_{max}	$< q_i >$	t^1_{PLM}	t^{20}_{PLM}
no	10^3	2511	4	10	39	8	$9\,10^2$	1306	4	10	$1.8\,10^4$	$1.5\,10^3$
0	10^3	2320	4	8.6	32	8	$9\,10^2$	1306	4	9.9	$1.8\,10^4$	$1.5\,10^3$
0.01	10^3	2065	4	5.8	22	7	$9\,10^2$	1305	4	5.8	10^4	10^3
0.1	$8\,10^2$	1350	4	2.2	11	7	$6\,10^2$	1314	4	2.3	$2\,10^3$	$3\,10^2$

TABLE IV: Running times, in seconds, for spACE ($N_2^{max} = 4N$) and PLM on 1 and 20 cores, for the data from the reference graph ER05, at 2 different sampling depth ER 1000 ($N = 50B = 1000$), ER100000 (N=50,B=100000), for different color compression thresholds f_0. The number of clusters processed by the algorithm $\hat{N}_{cl}$ and the average number of colors explicitly modeled are also indicated $\langle q_{kept} \rangle$.

	WW						PDZ						RRM1					
$f0$	t_{ACE}	$\hat{N}_{cl}$	k_{max}	$< q_i >$	t^1_{PLM}	t^{20}_{PLM}	t_{ACE}	$\hat{N}_{cl}$	k_{max}	$< q_i >$	t^1_{PLM}	t^{20}_{PLM}	t_{ACE}	$\hat{N}_{cl}$	k_{max}	$< q_i >$	t^1_{PLM}	t^{20}_{PLM}
no	$1.5\,10^3$	702	4	21	$2\,10^2$	25	$4\,10^4$	4355	5	21	10^4	838	$2\,10^5$	4127	6	21	$6\,10^4$	$4\,10^3$
0	10^3	702	4	18.6	$2\,10^2$	24	$3\,10^4$	4355	5	19.8	10^4	728	$2\,10^5$	4127	6	20.9	$6\,10^4$	$4\,10^3$
0.01	$4\,10^2$	702	4	9.4	85	15	$5\,10^3$	4548	6	11.2	$5\,10^3$	428	$6\,10^4$	4992	7	12	$3\,10^4$	$2\,10^3$
0.1	$2.5\,10^2$	714	4	2.2	17	8	10^3	4146	5	2.5	$1.5\,10^3$	123	10^3	3724	5	2.5	$5\,10^3$	$5\,10^2$

TABLE V: Running times (in seconds) for ACE and PLM for the three studied protein families: WW (N=31 B=8251), PDZ (N=84,B=24795), RRM1 (N=82,B=70780) for different color compression thresholds f_0. spACE has been stopped at $N_2^{max} = 93,252,246$ (WW, PDZ, RRM1); the number of clusters processed by the algorithm $\hat{N}_{cl}$ and the average number of colors explicitly modeled are also indicated $< q_{kept} >$. PLM times are given for both on 1 and 20 cores.

.

computation of the partition function which grows as $< q >^K$ on clusters of size K, the overall number of clusters which enters in the construction rule to and the number of configurations sampled by Monte-Carlo to calculate the relative errors in the reconstruction of the statistics of the data.
For PLM the average over the sampled configurations for the pseudo-likelihood and moments calculation determines a linear dependence on the sample size B, and on the number of parameters. Fig. 10 and Table IV show the computational times for the reference ER graph realization as a function of f_0 and B. On such data spACE (Fig. 10) weakly depends on K and $\langle q \rangle$,which take small values, the limiting step is therefore the number of Monte Carlo sampled configurations. Having a large number of MC steps (here fixed to 500000) is important to correctly estimate the relative errors at very large sample size, given the small value of the sampling variances. The computational time show therefore a weak linear dependence on the compression, as shown in Fig. 13 (Appendix). The computational times for spACE and PLM on real proteins are compared in Tables V and in Fig. 11. As shown in Table V for real proteins a larger numbers and larger sizes of clusters are summed on in the ACE expansion, as compared to the artificial data in Table IV. The total number of processed clusters ($\hat{N}_{cl}$) and their average size $\langle q \rangle$ are clearly the limiting factor for the running time, making ACE slower than PLM. PLM further benefits from parallel computing, here results obtained on 1 and 20 are compared, while for spACE the computation has also been followed by time consuming MC-learning. Fig. 11 shows the large reduction in computational time within spACE by color compression, such compression makes often feasible the numerical computation, especially for non sparse ACE inference (see Fig. 13) which would take prohibitively long times without significant compression.

VII. CONCLUSION

The present work reports an extensive numerical benchmarking of inferred color compressed Potts model with 2 algorithms (ACE and PLM) on synthetic data, as well as on protein sequence data. Knowing the

ground-truth model that generated the data, we can assess the inference performance at different compression strengths, by (1) computing the Kullback-Leibler divergence between the real and inferred models; (2) checking the reconstruction of low-order statistics; (3) testing the reconstruction of the structure of the interaction network, and of the couplings and field parameters.
We will in the following resume and discuss the comparison between PLM and ACE inference methods before discussing the results on color compression.
A first important advantage shown by ACE with respect to PLM inference is that ACE can be easily stopped at large value of the cluster inclusion threshold to reconstruct a good sparse model for the artificial data, which have been generated from sparse random graphs.Imposing sparsity regularization naturally reduces the number of coupling parameters in the inferred model, leaving the possibility to choose a small value for the L_2 regularization parameter on the left non-zero couplings. Such double regularization scheme allows a very good reconstruction of the overall model, the model parameters and the statistics. On the contrary for PLM , the inferred model is fully connected and therefore a large L_2 regularization is needed to avoid overfitting, in the optimal model reconstruction, with the consequences that all the amplitudes of coupling parameters are systematically underestimated, and the statistics of the data less well reproduced. Further theoretical investigations are needed to obtain the optimal regularization value found for the inference of a fully connected model, as a function of the sparsity of the original graph, the number of sites and color and the number of data and will be carried out on a forthcoming paper [49] . It would be also interesting to introduce a double regularization scheme for PLM algorithm. The situation is different on protein sequence data because the spACE procedure with the simple stopping criterium implemented here does not converge to small statistical reconstruction errors, without additional MC learning algorithm. The resulting inferred network is no more sparse and no gain of performances of ACE with respect to PLM are achieved at larger computational costs. spACE procedure can be improved to better impose sparsity constraint in a more general case [50, 51]. A question which deserves further investigation is if the inference of sparse connectivity graph is more appropriate for parameter reconstruction in the large undersampling regime, even for data generated with models, such as protein sequence data, which are not necessary sparse. Preliminary results for fitness prediction on protein sequence data seems to indicate that sparse models gives very good performances [51]. Next we will resume conclusion and discussion on color compression.
The central finding of the present work is that color compression does not degrade the studied performances in a very large range of frequency compression cut off, while largely reducing the dimensionality of the inferred model and its computational time. The reduction of computational time obtained thanks to color compression often become essential for solving the inverse problem in reasonable times. For example, ACE inference of models for many HIV proteins [52–55] has been possible thanks to color compression. The color compressed version of the PLM algorithm introduced adapted here from the routine of the group of Aurell and collaborators [9, 28] can be crucially important when dealing with large protein sequences or with whole genome inference [27, 56] with a much larger number of variables than single protein sequence data. The color compression and decompression procedures introduced here are not restricted to pairwise graphical models. They could be used in other machine learning approaches, on protein sequence data, such as restricted Botzmann machines [57], or variational auto-encoders [58] .

Acknowledgements. We thank Lorenzo Posani for useful discussions. This work benefited from the financial support of the grants PSL ProTheoMicS and RBMPro ANR-17-CE30-0021-01.

VIII. APPENDIX

A. Reminder about Adaptive Cluster Expansion and the inclusion threshold

In the ACE inference procedure the cross entropy is expanded as the sum of cluster contributions. Defining a cluster as a sub-set of variables: $\Gamma = \{i_1, \ldots, i_k\}, k \leq N$, we can formally write the cross-entropy as the sum of cluster contributions:

$$S(\mathbf{J}|\mathbf{f}) = \sum_{\Gamma} \Delta S_{\Gamma}, \tag{18}$$

where the sum is over all nonempty clusters of the N variables. The cluster cross-entropy contributions ΔS_Γ are recursively defined through

$$\Delta S_\Gamma = S_\Gamma - \sum_{\Gamma' \subset \Gamma} \Delta S_{\Gamma'} \,. \tag{19}$$

Here S_Γ denotes the minimum of the cross entropy (4) restricted only to the variables in Γ. Thus, S_Γ depends only on the frequencies $p_i(a)$, $p_{ij}(a,b)$ with $i,j \in \Gamma$. Provided that the number of variables in Γ is small (typically $\lesssim 10$ for $q = 10$ Potts state as in the present work) numerical maximization of the likelihood restricted to Γ is tractable. The definition of ΔS_Γ ensures that the sum over all clusters Γ in (18) yields the cross entropy for the entire system of N variables. As detailed in [2, 32], a recursive construction rule is used to avoid, before selection, the computation of all cluster entropies. Such rule consists in building up clusters of size k by combining selected clusters selected of size $k-1$. The ACE expansion consists in truncating the expansion in Eq. (18) by fixing a cluster inclusion threshold t and summing up in Eq.(19) only cluster contribution with $|\Delta S_\Gamma| > t$.

B. ACE and SpACE pseudocodes

The ACE algorithm, described in detail in [32], is based on the selection and summation of individual cluster contributions to the Cross Entropy. It is built by the recursion of the following routine. Given a list L_k of clusters of size k, beginning with the list of all the $N\ (N-1)/2$ possible clusters of size $k = 2$:

1. For each cluster $\Gamma \in L_k$
   (a) Compute S_Γ by numerical minimization of Eq. (4) restricted to Γ.
   (b) Record the parameters minimizing Eq. (4), called $\mathbf{J}_\Gamma$.
   (c) Compute ΔS_Γ using Eq. (19).
2. Add all clusters $\Gamma \in L_k$ with $|\Delta S_\Gamma| > t$ to a new list $L'_k(t)$.
3. Construct a list L_{k+1} of clusters of size $k+1$ from overlapping clusters in $L'_k(t)$.

The rule used by default for constructing new clusters of size $k+1$ from selected clusters of size k is the so called strict rule: a new cluster is added only if all of its $k+1$ subclusters of size k belong to $L'_k(t)$. The above process is then repeated until no new clusters can be constructed (for ACE) or also until the maximal number of 2-site clusters N_2^{max} is in the selected list $L'_k(t)$ (for SpACE). After the summation of clusters terminates, the approximate value of the parameters minimizing the cross-entropy, given the current value of the threshold, is computed by

$$\mathbf{J}(t) = \sum_k \sum_{\Gamma \in L'_k(t)} \Delta \mathbf{J}_\Gamma, \quad \Delta \mathbf{J}_\Gamma = \mathbf{J}_\Gamma - \sum_{\Gamma' \subset \Gamma} \Delta \mathbf{J}_{\Gamma'} \,. \tag{20}$$

Then a Monte Carlo simulation is run to estimate the model one and two point correlations functions, which are compared to the empirical ones, taking into account their expected statistical fluctuations, by the computations of relative errors defined in Eq. (13, 14, 15) see also Fig. 2 in main text. When the maximal error is smaller than one the algorithm stops. Within the SpACE approximation even if a relative error of 1 is not reached, the smaller relative errors among the ones obtained for logarithmically spaced threshold intervals is chosen.

C. Cluster expansion and computational time as a function of the color compression f_0 for fully connected graphs.

Fig. 12 shows that the behavior of the maximal relative reconstruction error ϵ_{max} as a function of the cluster inclusion threshold t when changing the color compression threshold f_0. The presence of two relative minima corresponding to a sparse and a fully connected models fitting the data is observed for all the values

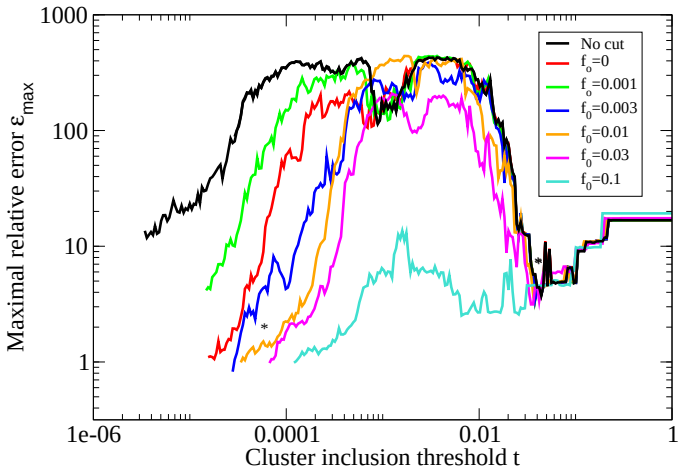


FIG. 12: Maximum relative error as a function of the expansion threshold for a particular graph realization (same used in Fig 5: ER05, sampled with B=1000),for different color compression f_0 .

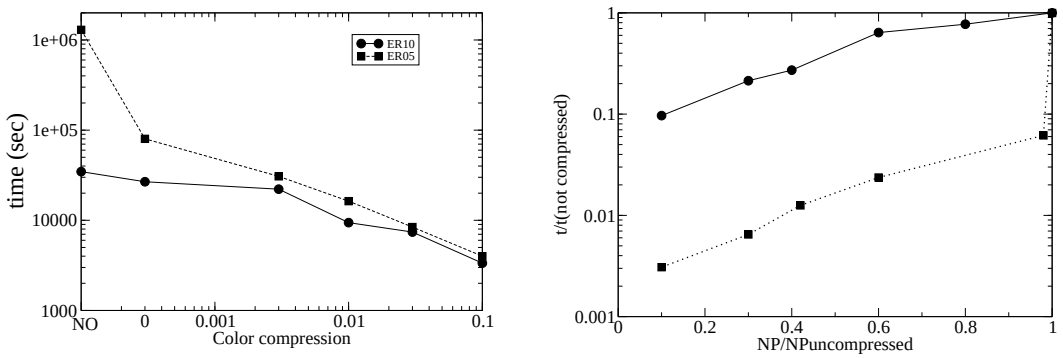


FIG. 13: Reduction in computational time due to the color compression for fully connected ACE inference on 2 data sets obtained by sampling B=1000 configurations from two Erdős-Rényi random graph models. Left: Computational time at the fully connected, low-threshold minimum, as a function of the color compression threshold f_0. Right: Computational time relative to the one with no color compression as a function of the number of parameters.

of f_0, see the two stars in Fig. 12. Moreover the threshold t corresponding to the sparse inferred graph is largely independent of the level of color compression.

Fig. 10 in the main text shows a mild computational gain as a function of the color compression when inferring a sparse interaction network (large-threshold minima). Such gain is generally huge when the expansion converges only at low threshold values and sums up clusters of larger and larger sizes K. The numerical computation of the cross entropy requires indeed the sums over a number of q^K configurations for K Potts variables with q states each. To illustrate this effect in Fig. 13 we show the reduction in computational time when the cluster expansion is stopped at the small threshold value corresponding to a fully connected inferred graph. One can reach a 1000-fold computational time reduction with large color compressions. As shown in Fig. 13 the expansion was stopped to maximal relative error of order 10 at small threshold t after 11 days while it took 50 minutes to infer a good quality, fully connected model for the

maximal color compression $f_0 = 0.1$. Note that the computational time to reach the sparse good model shown in Fig. 10 is smaller due to the reduced number of clusters. For large interconnected models color compression can therefore be essential to reach convergence in a reasonable amount of time and infer a model that reproduces the statistics of the data.

D. Kullback-Leibler Divergence from the ACE expansion

The computation is done in the Ising case for the simplicity of the notations, the generalization to the Potts case being straightforward. We denote $\boldsymbol{J}^B = \{J_{ij}^B, h_i^B\}$ the inferred parameters at sample size B, and $\boldsymbol{J}^{true} = \{J_{ij}^{true}, h_i^{true}\}$ the true underlying model parameters. The inferred cross-entropy at sampling B writes

$$S_B = -\sum_{\boldsymbol{\sigma}} P_{\boldsymbol{J}^B}(\boldsymbol{\sigma}) \log P_{\boldsymbol{J}^B}(\boldsymbol{\sigma}) \ , \tag{21}$$

where the sum is over all possible configurations $\boldsymbol{\sigma} = \{\sigma_1, ..., \sigma_N\}$. The inferred probability distribution at finite sampling B is

$$P_{\boldsymbol{J}^B}(\boldsymbol{\sigma}) = \frac{\exp\left(\sum_{i=1}^N h_i^B \sigma_i + \sum_{k<l}^N J_{kl}^B \sigma_k \sigma_l\right)}{\mathcal{Z}_B} \ . \tag{22}$$

The Kullback-Leibler (KL) divergence between the true and the inferred distributions writes

$$\begin{aligned} D(P_{\boldsymbol{J}^{true}}||P_{\boldsymbol{J}^B}) &= \sum_{\boldsymbol{\sigma}} P_{\boldsymbol{J}^{true}}(\boldsymbol{\sigma}) \log \frac{P_{\boldsymbol{J}^{true}}(\boldsymbol{\sigma})}{P_{\boldsymbol{J}^B}(\boldsymbol{\sigma})} \\ &= -S_{true} - \sum_{\boldsymbol{\sigma}} P_{\boldsymbol{J}^{true}}(\boldsymbol{\sigma}) \left\{ \sum_i h_i^B \sigma_i + \sum_{k<l} J_{kl}^B \sigma_k \sigma_l - \log \mathcal{Z}^B \right\} \\ &= -S_{true} + \log \mathcal{Z}^B - \sum_{\boldsymbol{\sigma}} P_{\boldsymbol{J}^{true}}(\boldsymbol{\sigma}) \left\{ \sum_i h_i^B \sigma_i + \sum_{k<l} J_{kl}^B \sigma_k \sigma_l \right\} \ . \end{aligned}$$

However, Eqs. (21) & (22) give

$$\log \mathcal{Z}^B = S_B + \sum_{\boldsymbol{\sigma}} P_{\boldsymbol{J}^B}(\boldsymbol{\sigma}) \left\{ \sum_i h_i^B \sigma_i + \sum_{k<l} J_{kl}^B \sigma_k \sigma_l \right\} \ .$$

The KL divergence between the true and the inferred distributions then writes

$$\begin{aligned} D(P_{\boldsymbol{J}^{true}}||P_{\boldsymbol{J}^B}) =& (S_B - S_{true}) - \sum_{\boldsymbol{\sigma}} P_{\boldsymbol{J}^{true}}(\boldsymbol{\sigma}) \left\{ \sum_i h_i^B \sigma_i + \sum_{k<l} J_{kl}^B \sigma_k \sigma_l \right\} \\ &+ \sum_{\boldsymbol{\sigma}} P_{\boldsymbol{J}^B}(\boldsymbol{\sigma}) \left\{ \sum_i h_i^B \sigma_i + \sum_{k<l} J_{kl}^B \sigma_k \sigma_l \right\} \ . \end{aligned}$$

Moreover, a reasonable approximation is

$$S_{true} = -\sum_{\boldsymbol{\sigma}} P_{\boldsymbol{J}^{true}}(\boldsymbol{\sigma}) \log P_{\boldsymbol{J}^{true}}(\boldsymbol{\sigma}) \approx S_{B\to\infty} = -\sum_{\boldsymbol{\sigma}} P_{\boldsymbol{J}^{B\to\infty}}(\boldsymbol{\sigma}) \log P_{\boldsymbol{J}^{B\to\infty}}(\boldsymbol{\sigma}) \ , \tag{23}$$

because the true underlying parameters are recovered by the inference method in the perfect sampling case: $P_{\boldsymbol{J}^{B\to\infty}}(\boldsymbol{\sigma}) \to P_{\boldsymbol{J}^{true}}(\boldsymbol{\sigma})$. Therefore,

$$D(P_{\boldsymbol{J}^{true}}||P_{\boldsymbol{J}^B}) = (S_B - S_\infty) + \sum_i h_i^B (\langle \sigma_i \rangle^B - \langle \sigma_i \rangle^\infty) + \sum_{k<l} J_{kl}^B (\langle \sigma_k \sigma_l \rangle^B - \langle \sigma_k \sigma_l \rangle^\infty) \ , \tag{24}$$

where $\langle\cdot\rangle^B = \sum_{\boldsymbol{\sigma}} \cdot P_{\boldsymbol{J}^B}(\boldsymbol{\sigma})$, and $\langle\cdot\rangle^\infty = \sum_{\sigma} \cdot P_{\boldsymbol{J}^{B\to\infty}}(\boldsymbol{\sigma}) \approx \sum_{\sigma} \cdot P_{\boldsymbol{J}^{true}}(\boldsymbol{\sigma})$. It naturally generalizes to the q-states Potts case:

$$\begin{aligned} D(P_{\boldsymbol{J}^{true}}||P_{\boldsymbol{J}^B}) =& (S_B - S_\infty) + \sum_{i=1}^{N}\sum_{a=1}^{q} h_i^B(a)(\langle\sigma_{ia}\rangle^B - \langle\sigma_{ia}\rangle^\infty) \\ &+ \sum_{\substack{k,l=1\\k<l}}^{N}\sum_{c,d=1}^{q} J_{kl}^B(c,d)(\langle\sigma_{kc}\sigma_{ld}\rangle^B - \langle\sigma_{kc}\sigma_{ld}\rangle^\infty) \ . \end{aligned} \tag{25}$$

The artificial data are in a compressed representation (*cf.* Section III). The complete inferred parameters are recovered as explained in Eq. (12).

E. Assignment of fields to zero-frequency states after inference

In section III we have discussed the decompression method used in the paper. In particular, we have seen that a pseudo-count is associated to the unseen states to assign them a field with respect to the reference of the grouped/compressed state or the least probable state and, in principle, this is different to what implicitly done when the model is inferred without color compression. To better understand the difference between the two approaches let us consider a simplified example of an independent-site model. Without color compression, the fields are obtained as the minimum of:

$$S_{ind} = \log\sum_{a=1}^{q} e^{h_i(a)} - \sum_{a=1}^{q} h_i(a)\ p_i(a) + \gamma_h \cdot \sum_{a=1}^{q} h_i(a)^2 \tag{26}$$

which, for the unseen colors in the gauge of $Z = \sum_{a=1}^{q} e^{h_i(a)} = 1$, gives:

$$h_u^\Gamma = -Lw(\frac{1}{2\gamma_h}) \tag{27}$$

where $Lw(y)$ is the Lambert function, solution of $xe^x = y$. On the other hand, the field which we assign to these symbols during color decompression is, in the same gauge,

$$h_u^p = \log(\frac{\alpha}{B}) \tag{28}$$

where we set, as in the rest of the paper, $\alpha = 0.1$. In this independent-site approximation there is then a shift between the two procedures given by $\Delta h = h_i(a)^\Gamma - h_i(a)^p$, that depends from the pseudocount α and from the value of the regularization γ_h. In table VI we give these shifts for the two values of γ_h used respectively by ACE ($\gamma_h = 0.01/B$) and PLM ($\gamma_h = 0.1/B$).

If, in the approximation of independent-sites, the difference Δh is the same in all gauges, the specific values of $h_i(a)^\Gamma$ and $h_i(a)^p$ of table VI are specific of the $Z = 1$ gauge. To have a comparison in the consensus gauge as done in the rest of the paper one has to subtract $h_i(a)^\Gamma$ and $h_i(a)^p$ the field of the most common color c_i on the considered site. In the independent-site approximation this is just $h_i(c_i) = log(p_i(c_i))$, and makes that unseen colors of different sites are found to have different fields. In Fig. 14 we plot the fields for the unseen symbols with a color compression $f_0 = 0.01$ (green diamonds) and $f_0 = 0$ (blue squares) versus the one for no color-compression for PLM and ACE (same fields as in Fig. 7 of the main paper), in the consensus gauge. We can see a systematic shift towards lower values (at least for PLM, to be checked for ACE). We can compare this shift with the theoretical shift obtained with the independent model as described above. Even if we neglect the terms due to the couplings we can well reproduce such shift as the difference between the field obtained with the pseudo-count with respect to the one obtained with the regularization, there is a good agreement between what observed and the theoretical shift for independent variables. In particular the shift is smaller for the regularization chosen by the ACE procedure.

B	$h_u^{\Gamma(ACE)}$	h_u^p	Δh_u	B	$h_u^{\Gamma(PLM)}$	h_u^p	Δh_u
10^2	-6.6	-6.9	0.3	10^2	-4.7	-6.9	1.5
10^3	-8.6	- 9.2	0.5	10^3	-6.6	-9.2	2.5
10^4	-10.7	-11.5	0.8	10^4	-8.7	-11.5	2.8
10^5	-12.9	-13.8	0.9	10^5	-10.7	-13.8	3.1

TABLE VI: Difference between the fields fixed by regularization and the one computed with the pseudo-count for the unseen Potts variables in the approximation of independent sites respectively for the fields regularization $\gamma_h = 0.01/B$ used in ACE and $\gamma_h = 0.1/B$ used in PLM

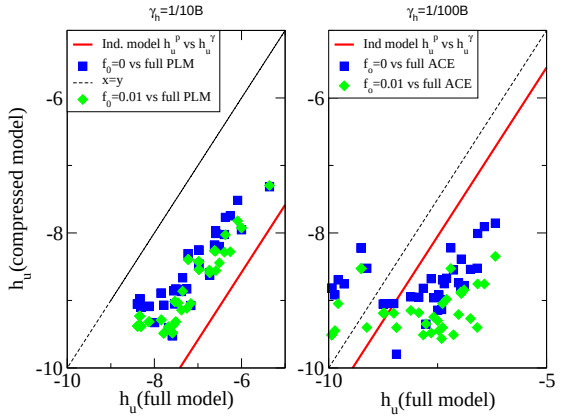


FIG. 14: Fields for the unseen Potts symbols for ER05 and $B = 1000$ in the non compressed model and the reference model with $f_0 = 0.01$ of Fig. 7. Dotted line x=y. Red Line: the shift given in the independent model for $\gamma_h = 1/10B$ as the one used with PLM.

1. *KL divergence for PLM at low L_2 regularization as a function of the sampling depth B and of the color compression threshold f_0*

In this section we study what happens with PLM at lower L_2 regularization, *e.g.* $\gamma_J = 1/B$ as the one used for ACE, when the threshold for color compression is varied. Without color compression, the performance obtained for $\gamma_J = 1/B$ becomes significantly worse, see Table III.

Figure 15 shows the average KL divergence between the true model and the inferred one for several color compression frequencies at low L_2 regularization ($\gamma_J = 1/B$). For all the sample sizes, we observe the existence of an optimal value of f_0. Especially at small sampling depth $B \leq 1000$, large color compression leads to a very significant decrease in the KL divergence. However, in spite of the improvement due to color compression, the KL divergences do not reach the minimal values obtained with strong L_2 regularizations (dashed lines in Fig. 15), showing that a good choice of the regularization is always essential.

F. Graphical reconstruction: Contact maps and *Fscore* for spACE graph reconstruction as a function of color compression

In Figs. 16 we compare the real contact maps with the heat map of the Frobenius norms of the couplings inferred by ACE and PLM with no color compression and $f_0 = 0.01$ for the reference data (ER05, $B = 1000$). Results at different color compressions are very similar, such as the identification of largely coupled sites by ACE and PLM.

To encompass both the precision and the recall in a single measure it is possible to use the Fscore, which

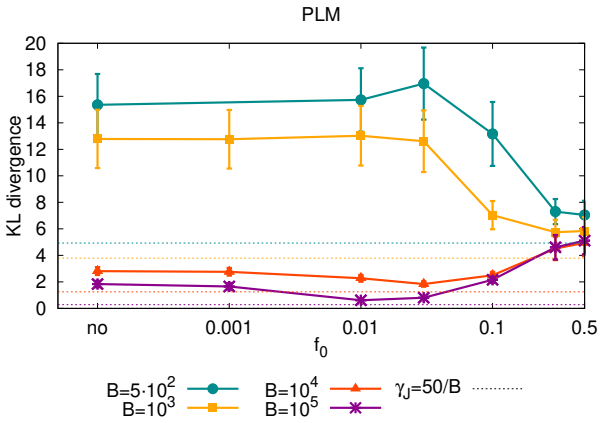


FIG. 15: KL divergence between the true model and the one inferred with $\gamma_J = 1/B$ averaged over 10 realizations for several sample sizes B at different color compression thresholds f_0 (full line). Error bars are standard deviations over the 10 realizations. Horizontal dashed lines are there for comparison and correspond to the KL values obtained with $\gamma_J = 50/B$ without color compression.

is the harmonic mean between the two and gives

$$\text{Fscore} = 2\frac{TP(N_{pred})}{N_{pred} + N_0} \tag{29}$$

where TP is the number of true predicted contacts, N_{pred} is the number of predicted contacts and N_0 is the number real contacts. The Fscore for ACE inference are plotted in Fig.17.

G. Definition of Positive Predicted Value and Average Product Correction

The Positive Predicted Value (PPV) curve is defined as:

$$PPV(n) = \frac{TP(n)}{n}. \tag{30}$$

Where $TP(n)$ is the number of true predicted edges in the top n pairs. the Average Product Correction (APC) [18, 23, 43],

$$F_{ij}^{APC} = F_{ij} - \frac{F_{i.}\,F_{.j}}{F_{..}}\,. \tag{31}$$

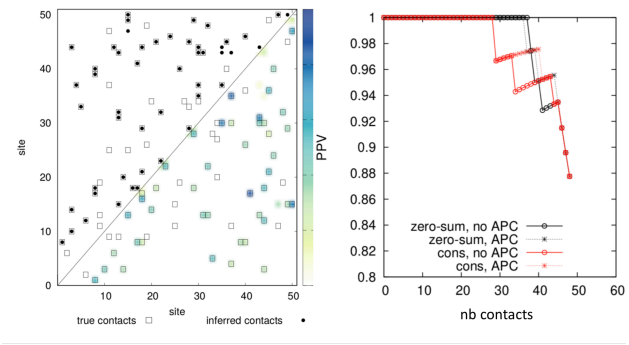
Here the dot indicates the average over the corresponding variables, e.g. $F_{i,.}$ is the average of F_{ij} over the second index j.

H. Reweighting procedure

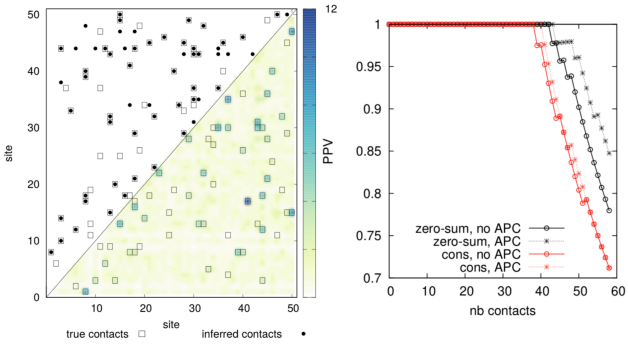
To reduce sampling bias, we decrease the statistical weight of sequences having many similar ones in the MSA. More precisely, the weight of each sequence is defined as the inverse number of sequences within Hamming distance $d_H < xL$, with an arbitrary but fixed $x \in (0, 1)$:

$$w_m = \frac{1}{||\{n|1 \leq n \leq M; d_H[(a_1^n, ..., a_L^n), (a_1^m, ..., a_L^m)] \leq xL\}||} \tag{32}$$

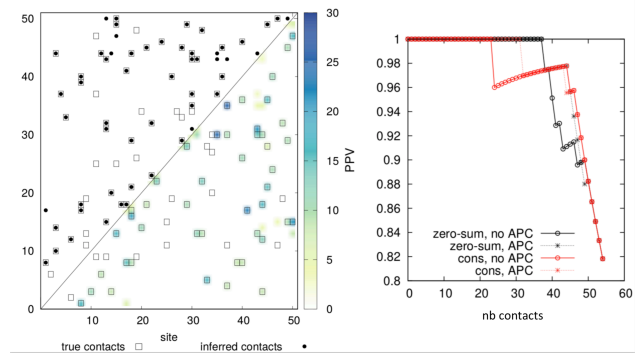
for all $m = 1, ..., M$. The weight equals one for isolated sequences, and becomes smaller the denser the sampling around a sequence is. Note that $x = 0$ would account to removing double counts from the MSA.



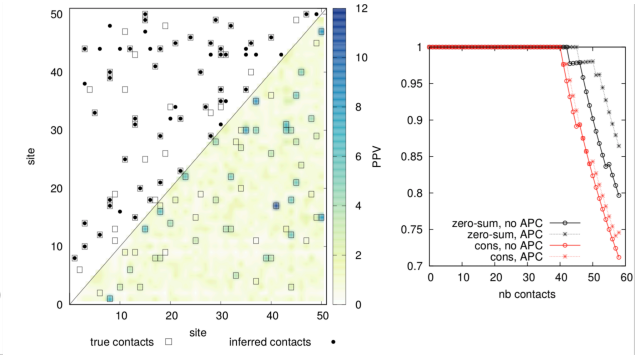
(a) ACE: no color compression



(b) PLM: no color compression



(c) ACE: $f_0 = 0.01$



(d) PLM: $f_0 = 0.01$

FIG. 16: ACE and PLM interaction graph reconstruction and PPV curve for one realization of ER graph. Left: contact maps. Upper triangular: contact map with real contacts (black empty squares), inferred true positive (full circles in empty squares), inferred false positive (full circles) in consensus gauge without Average Product Correction (APC). Lower triangular: Frobenius norm of the inferred parameters in consensus gauge with color-scale on the right. Right: Positive Predicted Value (PPV) curve in consensus and zero-sum gauge with and without APC. Top: no color compression. Bottom: $f_0 = 0.01$.

The total weight

$$M_{eff} = \sum_{m=1}^{M} w_m \tag{33}$$

can be interpreted as the effective number of independent sequences.

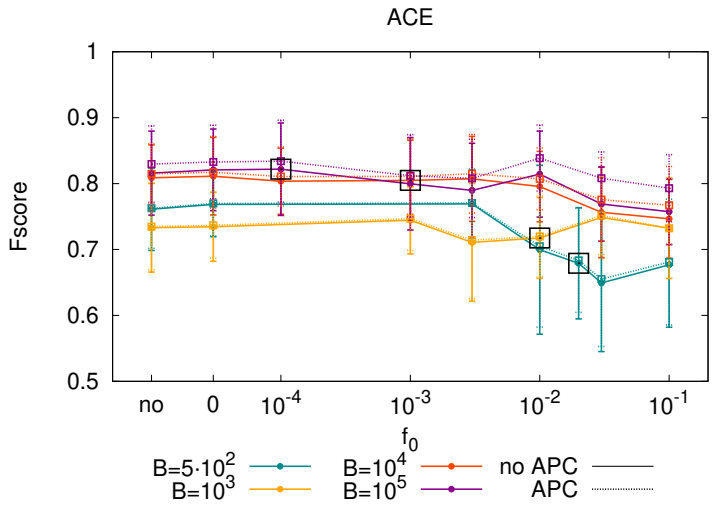


FIG. 17: Fscore for ACE inference. Points and error bars are averages and standard deviations obtained on 10 ER realizations. All quantities are computed in the consensus gauge with (dotted line) and without (full line) APC.

[1] D. H. Ackley, G. E. Hinton, and T. J. Sejnowski, "A learning algorithm for boltzmann machines," *Cognitive science*, vol. 9, no. 1, pp. 147–169, 1985.

[2] S. Cocco and R. Monasson, "Adaptive cluster expansion for the inverse ising problem: convergence, algorithm and tests," *Journal of Statistical Physics*, vol. 147, no. 2, pp. 252–314, 2012.

[3] H. C. Nguyen, R. Zecchina, and J. Berg, "Inverse statistical problems: from the inverse ising problem to data science," *Advances in Physics*, vol. 66, no. 3, pp. 197–261, 2017.

[4] F.-Y. Wu, "The potts model," *Reviews of modern physics*, vol. 54, no. 1, p. 235, 1982.

[5] S. Cocco, R. Monasson, L. Posani, and G. Tavoni, "Functional networks from inverse modeling of neural population activity," *Current Opinion in Systems Biology*, vol. 3, pp. 103–110, 2017.

[6] E. Schneidman, M. J. Berry, R. Segev, and W. Bialek, "Weak pairwise correlations imply strongly correlated network states in a neural population," *Nature*, vol. 440, no. 7087, pp. 1007–1012, 2006.

[7] S. Cocco, S. Leibler, and R. Monasson, "Neuronal couplings between retinal ganglion cells inferred by efficient inverse statistical physics methods," *Proceedings of the National Academy of Sciences*, vol. 106, no. 33, pp. 14058–14062, 2009.

[8] W. Bialek, A. Cavagna, I. Giardina, T. Mora, E. Silvestri, M. Viale, and A. M. Walczak, "Statistical mechanics for natural flocks of birds," *Proceedings of the National Academy of Sciences*, vol. 109, no. 13, pp. 4786–4791, 2012.

[9] M. Ekeberg, C. Lövkvist, Y. Lan, M. Weigt, and E. Aurell, "Improved contact prediction in proteins: using pseudolikelihoods to infer potts models," *Physical Review E*, vol. 87, no. 1, p. 012707, 2013.

[10] L. Burger and E. Van Nimwegen, "Disentangling direct from indirect co-evolution of residues in protein alignments," *PLoS Comput Biol*, vol. 6, no. 1, p. e1000633, 2010.

[11] M. Weigt, R. A. White, H. Szurmant, J. A. Hoch, and T. Hwa, "Identification of direct residue contacts in protein–protein interaction by message passing," *Proceedings of the National Academy of Sciences*, vol. 106, no. 1, pp. 67–72, 2009.

[12] D. S. Marks, L. J. Colwell, R. Sheridan, T. A. Hopf, A. Pagnani, R. Zecchina, and C. Sander, "Protein 3d structure computed from evolutionary sequence variation," *PloS one*, vol. 6, no. 12, p. e28766, 2011.

[13] J. I. Sułkowska, F. Morcos, M. Weigt, T. Hwa, and J. N. Onuchic, "Genomics-aided structure prediction," *Proceedings of the National Academy of Sciences*, vol. 109, no. 26, pp. 10340–10345, 2012.

[14] T. A. Hopf, L. J. Colwell, R. Sheridan, B. Rost, C. Sander, and D. S. Marks, "Three-dimensional structures of membrane proteins from genomic sequencing," *Cell*, vol. 149, no. 7, pp. 1607–1621, 2012.

[15] T. Nugent and D. T. Jones, "Accurate de novo structure prediction of large transmembrane protein domains using fragment-assembly and correlated mutation analysis," *Proceedings of the National Academy of Sciences*, vol. 109, no. 24, pp. E1540–E1547, 2012.

[16] D. de Juan, F. Pazos, and A. Valencia, "Emerging methods in protein co-evolution," *Nature Reviews Genetics*, vol. 14, no. 4, pp. 249–261, 2013.

[17] D. T. Jones, D. W. Buchan, D. Cozzetto, and M. Pontil, "Psicov: precise structural contact prediction using sparse inverse covariance estimation on large multiple sequence alignments," *Bioinformatics*, vol. 28, no. 2, pp. 184–190, 2012.

[18] F. Morcos, N. P. Schafer, R. R. Cheng, J. N. Onuchic, and P. G. Wolynes, "Coevolutionary information, protein folding landscapes, and the thermodynamics of natural selection," *Proceedings of the National Academy of Sciences*, vol. 111, no. 34, pp. 12408–12413, 2014.

[19] A. L. Ferguson, J. K. Mann, S. Omarjee, T. Ndungu, B. D. Walker, and A. K. Chakraborty, "Translating hiv sequences into quantitative fitness landscapes predicts viral vulnerabilities for rational immunogen design," *Immunity*, vol. 38, no. 3, pp. 606–617, 2013.

[20] J. K. Mann, J. P. Barton, A. L. Ferguson, S. Omarjee, B. D. Walker, A. Chakraborty, and T. Ndung'u, "The fitness landscape of hiv-1 gag: advanced modeling approaches and validation of model predictions by in vitro testing," *PLoS Comput Biol*, vol. 10, no. 8, p. e1003776, 2014.

[21] M. Figliuzzi, H. Jacquier, A. Schug, O. Tenaillon, and M. Weigt, "Coevolutionary landscape inference and the context-dependence of mutations in beta-lactamase tem-1," *Molecular biology and evolution*, vol. 33, no. 1, pp. 268–280, 2016.

[22] T. A. Hopf, J. B. Ingraham, F. J. Poelwijk, C. P. Schärfe, M. Springer, C. Sander, and D. S. Marks, "Mutation effects predicted from sequence co-variation," *Nature biotechnology*, vol. 35, no. 2, p. 128, 2017.

[23] S. Cocco, C. Feinauer, M. Figliuzzi, R. Monasson, and M. Weigt, "Inverse statistical physics of protein sequences: a key issues review," *Reports on Progress in Physics*, vol. 81, no. 3, p. 032601, 2018.

[24] M. Socolich, S. W. Lockless, W. P. Russ, H. Lee, K. H. Gardner, and R. Ranganathan, "Evolutionary information for specifying a protein fold," *Nature*, vol. 437, no. 7058, pp. 512–518, 2005.

[25] W. P. Russ, M. Figliuzzi, C. Stocker, P. Barrat-Charlaix, M. Socolich, P. Kast, D. Hilvert, R. Monasson, S. Cocco, M. Weigt, and R. Ranganathan, "Evolution-based design of chorismate mutase enzymes," *In preparation.*

[26] C. Baldassi, M. Zamparo, C. Feinauer, A. Procaccini, R. Zecchina, M. Weigt, and A. Pagnani, "Fast and accurate multivariate gaussian modeling of protein families: predicting residue contacts and protein-interaction partners," *PloS one*, vol. 9, no. 3, p. e92721, 2014.

[27] J. Pensar, Y. Xu, S. Puranen, M. Pesonen, Y. Kabashima, and J. Corander, "High-dimensional structure learning of binary pairwise markov networks: A comparative numerical study," *arXiv preprint arXiv:1901.04345*, 2019.

[28] M. Ekeberg, T. Hartonen, and E. Aurell, "Fast pseudolikelihood maximization for direct-coupling analysis of protein structure from many homologous amino-acid sequences," *Journal of Computational Physics*, vol. 276, pp. 341–356, 2014.

[29] M. Vuffray, S. Misra, and A. Y. Lokhov, "Efficient learning of discrete graphical models," *arXiv preprint arXiv:1902.00600*, 2019.

[30] J. Sohl-Dickstein, P. B. Battaglino, and M. R. DeWeese, "New method for parameter estimation in probabilistic models: minimum probability flow," *Physical review letters*, vol. 107, no. 22, p. 220601, 2011.

[31] S. Cocco and R. Monasson, "Adaptive cluster expansion for inferring boltzmann machines with noisy data," *Physical Review Letters*, vol. 106, no. 9, p. 090601, 2011.

[32] J. P. Barton, E. De Leonardis, A. Coucke, and S. Cocco, "Ace: adaptive cluster expansion for maximum entropy graphical model inference," *Bioinformatics*, vol. 32, no. 20, pp. 3089–3097, 2016.

[33] A. Haldane, W. F. Flynn, P. He, and R. M. Levy, "Coevolutionary landscape of kinase family proteins: sequence probabilities and functional motifs," *Biophysical journal*, vol. 114, no. 1, pp. 21–31, 2018.

[34] B. Anton, M. Besalu, O. Fornes, J. Bonet, G. De las Cuevas, N. Fernandez-Fuentes, and B. Oliva, "Radi (reduced alphabet direct information): Improving execution time for direct-coupling analysis," *bioRxiv*, p. 406603, 2018.

[35] E. T. Jaynes, "On the rationale of maximum-entropy methods," *Proceedings of the IEEE*, vol. 70, no. 9, pp. 939–952, 1982.

[36] A. Gelman, A. Jakulin, M. G. Pittau, and Y.-S. Su, "A weakly informative default prior distribution for logistic and other regression models," *The Annals of Applied Statistics*, pp. 1360–1383, 2008.

[37] J. P. Barton, S. Cocco, E. De Leonardis, and R. Monasson, "Large pseudocounts and l 2-norm penalties are necessary for the mean-field inference of ising and potts models," *Physical Review E*, vol. 90, no. 1, p. 012132, 2014.

[38] The first interval is $I/\tau < t < 1$, the second is for $1/\tau^2 < t < 1/\tau$ and so on. We have chosen here τ =3.4 (see command -trec on the GitHub site).

[39] R. Salakhutdinov, "Learning and evaluating boltzmann machines," tech. rep., UTML TR 2008?002, 2008.

[40] We estimate the standard deviation σ of the osberved frequency based on a Binomial variable with probability $\hat{f}$ in a sample of size B: $\sigma = \sqrt{(\hat{f}\,(1-\hat{f})/B} \simeq \sqrt{\hat{f}/B}$. Probabilities larger than f_0 correspond to signal-to-noise ratios $\hat{f}/\sigma > \sqrt{f_0^*\,B} > 3$.

[41] Similar results are obtained when considering the Pearson correlations between the real and inferred frequencies or correlations rather than the absolute errors (not shown).

[42] F. Morcos, A. Pagnani, B. Lunt, A. Bertolino, D. S. Marks, C. Sander, R. Zecchina, J. N. Onuchic, T. Hwa, and M. Weigt, "Direct-coupling analysis of residue coevolution captures native contacts across many protein families," *Proceedings of the National Academy of Sciences*, vol. 108, no. 49, pp. E1293–E1301, 2011.

[43] S. D. Dunn, L. M. Wahl, and G. B. Gloor, "Mutual information without the influence of phylogeny or entropy dramatically improves residue contact prediction," *Bioinformatics*, vol. 24, no. 3, pp. 333–340, 2008.

[44] J. Cardy, *Scaling and renormalization in statistical physics*, vol. 5. Cambridge university press, 1996.

[45] Similar results are obtained when considering the Pearson correlations between real and inferred parameters, rather than their absolute differences. ACE gives better results than PLM for parameter reconstruction, especially on couplings (not shown). This is due to the fact that spACE avoid overfitting of data and setting many non-zero couplings for non interacting sites in the real interaction graph.

[46] C. L. Araya, D. M. Fowler, W. Chen, I. Muniez, J. W. Kelly, and S. Fields, "A fundamental protein property, thermodynamic stability, revealed solely from large-scale measurements of protein function," *Proceedings of the National Academy of Sciences*, vol. 109, no. 42, pp. 16858–16863, 2012.

[47] R. N. McLaughlin Jr, F. J. Poelwijk, A. Raman, W. S. Gosal, and R. Ranganathan, "The spatial architecture of protein function and adaptation," *Nature*, vol. 491, no. 7422, p. 138, 2012.

[48] D. Melamed, D. L. Young, C. E. Gamble, C. R. Miller, and S. Fields, "Deep mutational scanning of an rrm domain of the saccharomyces cerevisiae poly (a)-binding protein," *Rna*, vol. 19, no. 11, pp. 1537–1551, 2013.

[49] A. Fanthomme, S. Cocco, and R. Monasson, "Optimal regularizations for multivariate gaussian distributions," *In Preparation*, 2019.

[50] S. Franz, F. Ricci-Tersenghi, and J. Rocchi, "A fast and accurate algorithm for inferring sparse ising models via parameters activation to maximize the pseudo-likelihood," *arXiv preprint arXiv:1901.11325*, 2019.

[51] S. Cocco, L. Posani, and R. Monasson, "Functional couplings from sequence and mutational data," *In Preparation*, 2019.

[52] J. P. Barton, M. Kardar, and A. K. Chakraborty, "Scaling laws describe memories of host–pathogen riposte in the hiv population," *Proceedings of the National Academy of Sciences*, vol. 112, no. 7, pp. 1965–1970, 2015.

[53] J. P. Barton, N. Goonetilleke, T. C. Butler, B. D. Walker, A. J. McMichael, and A. K. Chakraborty, "Relative rate and location of intra-host hiv evolution to evade cellular immunity are predictable," *Nature communications*, vol. 7, p. 11660, 2016.

[54] J. P. Barton, A. K. Chakraborty, S. Cocco, H. Jacquin, and R. Monasson, "On the entropy of protein families," *Journal of Statistical Physics*, vol. 162, no. 5, pp. 1267–1293, 2016.

[55] R. H. Louie, K. J. Kaczorowski, J. P. Barton, A. K. Chakraborty, and M. R. McKay, "Fitness landscape of the human immunodeficiency virus envelope protein that is targeted by antibodies," *Proceedings of the National Academy of Sciences*, vol. 115, no. 4, pp. E564–E573, 2018.

[56] C.-Y. Gao, H.-J. Zhou, and E. Aurell, "Correlation-compressed direct-coupling analysis," *Physical Review E*, vol. 98, no. 3, p. 032407, 2018.

[57] J. Tubiana, S. Cocco, and R. Monasson, "Learning protein constitutive motifs from sequence data," *eLife*, vol. 8, p. e39397, 2019.

[58] A. J. Riesselman, J. B. Ingraham, and D. S. Marks, "Deep generative models of genetic variation capture the effects of mutations," *Nat. Methods*, vol. 15, pp. 816–822, 2018.