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, with a non-uniform number of Potts colors (symbols) across 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 colors available to each variable is reduced, and interaction networks are made sparse. To achieve this color compression scheme, 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, the Adaptive Cluster Expansion (ACE) and the PseudoLikelihood Maximization (PLM) on synthetic data obtained by sampling disordered Potts models on an Erdős-Rényi random graphs. As expected inference with sparsity requirements outperforms inference of fully connected models (by ACE or PLM) when few data are available. Moreover we show 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.

[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

I. INTRODUCTION

Extracting patterns and information from vast databases has become one of the biggest challenges for scientists of many domains. Together with other machine-learning techniques, graphical models, are adequate tools to infer effective couplings between variables from data in many disciplines. 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-categorical variables[4]. Applications include inferring functional couplings among a set of neurons from their neural activity recording [5–7], dynamical couplings birds in flocks[8] and inferring couplings from collections of protein sequences that belong to the same homologous family [9]. PFAM [10, 11], for example, is a huge database providing protein sequences already aligned and organized by protein family. Over the last decade, it was shown that describing these protein families by Potts models, whose parameters are learned from the corresponding sequence alignment may provide information on the protein structure [9, 12–20], predict fitness variations following mutations [3, 21–24] and design new working proteins of the same family [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 [12, 19, 26, 27], message passing [13], PseudoLikelihood Maximization[9, 28], minimum probability flow [29], and the Adaptive Cluster Expansion (ACE) method [30, 31].

Even if modeling protein families as Potts models only approximates protein site interactions to, at most, pairwise interactions, the number of parameters to be inferred is still huge. For an N-site protein, where each site can be one 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-based regularization, on the contrary, may effectively remove many interaction parameters associated to low (in absolute value) connected correlations. However, poor sampling generally leads to very unreliable estimates of the correlations, which may take large values, and make L_1-regularization ineffective.

In this paper we propose a simple procedure to reduce the number of Potts parameters. We infer a compressed Potts model, where the number of states $q_i \leq q$ depends on the site i. The main idea is to group together rarely observed states on each site, defined as those below a given 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 $\sum_i^N q_i + \sum_{i<j}^N q_i\, q_j$. Slightly different schemes are based on grouping colors according to their entropy contributions to the site variability [31] or to their mutual information [32] or compression to a fixed number of colors [33], and comes to a similar outcome. This 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 only fixed by regularization. We also introduce and describe a procedure to recover, after inference, a full model with parameters for all the possible q states, which is needed to compare different color compressions and will be referred to as color decompression.

In physics Potts states are often referred to as colors, so we call this state reduction procedure *color compression*. Such color compression was already used by some of us in the ACE algorithm [31] 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.

We first benchmark the approach on some Potts models with quenched disorder defined on Erdős-Rényi (ER) random graphs, as described in section III A 2, and then present applications to protein data [25]. The first part of the paper briefly sketches the methodological background and the inference algorithms (Sec. II). The procedure of color compression and decompression is exposed in Section III A. We then assess the performances of the procedure on synthetic data generated from Potts model on random graphs in Section IV. Our study is carried out with two inference methods: the Adaptive Cluster Expansion (ACE), which was already implemented with color compression in [31], and PseudoLikelihood Maximization (PLM), whose implementation of color compression was developed for the purpose of the present comparison. We show that ACE can be straightforwardly forced to infer sparse interaction networks by stopping the expansion with large cluster inclusion threshold. We evaluate the quality of the models inferred at different compression levels and sparsity in terms of the Kullback-Leibler distance to the empirical distribution (Sec. IV A), of their

ability to reproduce the original low-order statistics (Sec. IV B), of the accuracy of reconstruction of the interaction network (Sec. IV C) and parameters (Sec. IV D), and of gain in computational effort (Sec. IV E). 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. Some conclusion and perspectives are presented in Sec. VI.

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, first of all the one-site and two-site frequencies $f_i(a)$, $f_{ij}(a, b)$. In general the Potts model defined above is the simplest, or maximum entropy [34], 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. The problem of finding the parameters $h_i(a)$, $J_{ij}(a, b)$ that satisfy Eq. 3 is referred to as the inverse Potts problem.

B. Cross-entropy and regularization

Formally, the inverse Potts problem is solved by the set of fields and couplings that maximize the average log-likelihood or, equivalently, those 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 even when starting with a finite data sample, a regularization term ΔS is added to the cross-entropy, which, in the Bayesian formulation, corresponds to a prior knowledge of the parameter distribution. A Gaussian prior distribution for the parameters, also referred to as L_2-regularization, is a typical 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. In the case that the regularization strengths are relatively weak ($\mathcal{O}(1/B)$), this regularization can be thought of as a weakly informative prior [35] 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 $f_{ij}(a,b)$, $i<j$ estimated from the data 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 this 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 a, b, $K_{ij}(a), 1 \le i,j \le N$, H_i and κ_{ij}. This freedom can be 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. In practice, fields and couplings can be simply put in the 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 with the description 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 VII A). The adaptive cluster expansion [2, 30, 31] 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 has been empirically shown in protein sequence analysis [9, 24, 25] and in theoretical analyses [36], 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. IV, and fixed $\gamma_J = 50/B$, $\gamma_h = 0.1/B$ for input sampling of size B.

With PLM gauge invariance is automatically broken. The inference is performed in the gauge that mini-

mizes the L_2-regularization:

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

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 depends on the site. This situation naturally arises due to sampling: states with very small probabilities are rarely observed. For instance, in multiple sequence alignments of real protein families, only a subset of the full $q = 21$ possible amino acids are observed for the large majority of sites. It also may arise as a result of our color compression procedure: for each site i, we model explicitly only the k_i states observed with a frequency $f_i(a)$ larger than the cutoff value

$$f_i(a) > f_0 \ , \tag{10}$$

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

2. *Illustration on Erdős-Rényi random graphs*

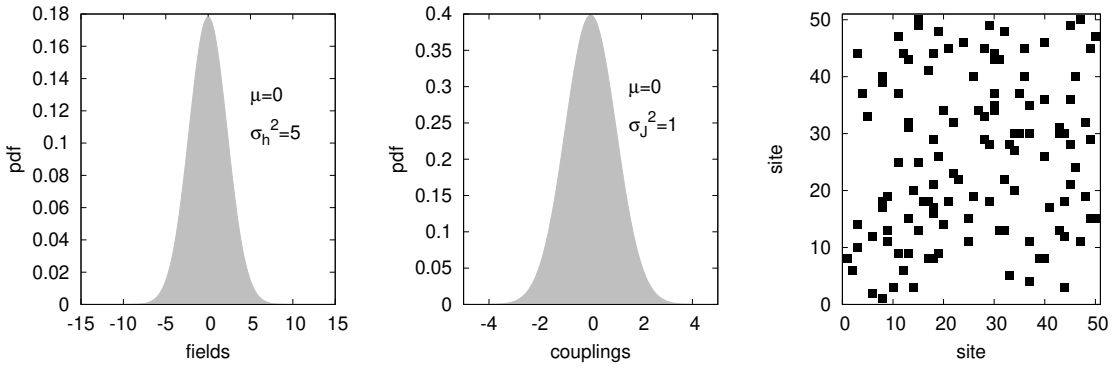


FIG. 1: ER models. Left and center: Gaussian distributions from which parameters of the ER models are chosen. For fields $\mu_h = 0$ and $\sigma_h^2 = 5$ (left) while for couplings $\mu_J = 0$ and $\sigma_J^2 = 1$ (center). Right: one particular realization of the interaction graph.

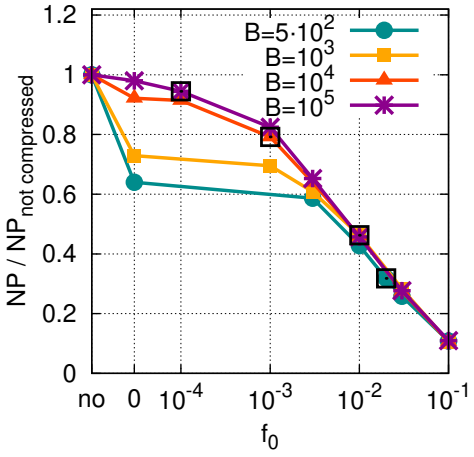


FIG. 2: Relative reduction in the number of parameters to be inferred as a function of the color compression for different sample sizes B on a single ER realization. f_0: no compression, 0 (only unseen symbols are removed from inference) and $1/B < f_0 < 0.1$. The number of parameters to be inferred without color compression is $NP_{decompressed} = 123000$.

To illustrate how color compression can reduce the number of variable states, we consider a Potts model with $q = 10$ states on an Erdős-Rényi random graph with $N = 50$ sites. Each edge in the network is included with probability 0.05 with a maximum connectivity equal to 7, and the Potts parameters on interacting sites are selected from Gaussian distributions of mean $\mu = 0$ and standard deviations $\sigma_J^2 = 1$ and $\sigma_h^2 = 5$ (Fig. 1). Regarding the states, no preferential scheme is imposed, i.e. if i and j interact, then 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.

We 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, $qN + q^2N(N-1)/2 \simeq 1.2\ 10^5$. 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.

Given the huge number of parameters that, a priori, characterizes our problem, an interesting question is how many of them are really important to fully describe the system, or more precisely, which parameters are really estimated from data and which are mostly determined by the regularization choice. Fig. 2 shows how the fraction of parameters effectively employed during the inference process scales when using color compression. It is evident that, already at small f_0, the initial number can be significantly reduced, especially at low sample size B. We will see in Sec. V that for real proteins, where $q = 21$ instead of 10, this phenomenon is even more pronounced.

3. *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 with the true one. To this aim we associate to the explicitly modeled states the same fields and couplings as in the reduced model. To determine the parameters for states that were *grouped* or to states that were never observed in the sampling, hereafter referred to as *unseen* states, we use the following procedure: For each grouped state a' the fields and the couplings are obtained as

$$\begin{aligned} h_i(a') &= h_i(k_i+1) + \log\left(\frac{f_i(a')}{f_i(k_i+1)}\right) , \\ J_{ij}(a',b) &= J_{ij}(k_i+1,b) . \end{aligned} \tag{11}$$

This procedure allows us to correctly recover the frequencies $f_i(a')$ even for the different grouped Potts states, while a common coupling parameters is assigned for all the grouped states.

Then, one needs to associate fields and couplings to the never observed states, on which no direct information is available. A natural extension to the procedure described above for the grouped states is to fix them in reference to the grouped state by Eq. 11. To this end we assign a pseudocount frequency $f'' = \alpha/B$ to these never observed states (a''). When the grouped state is not present we fix them in reference to the least probable state (a''') by using the same Eq. 11 with $k_i + 1$ replaced by a'''. In the results shown here we have fixed $\alpha = 0.1$, in the expected range $0 < \alpha < 1$. In appendix VII E we compare the role of the pseudo-count and the standard L_2 regularization by comparing the fields obtained by the procedure described above with the one obtained by the minimization of the cross entropy with L_2 regularization (Eqs. 4 and 5) in an independent model. 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.

4. *Gauge used in the ACE inference*

ACE inference is always done after removal of one Potts state, which defines a gauge for the field and coupling parameters. The ACE algorithm is also based on an expansion of the partition function around the Boltzmann weight for the gauge symbol, which is one, to speed up its calculation [31].

The choice of the symbol to remove 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 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 gauging to zero 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 gauged to zero, and have therefore null statistical variances by construction. Fields and couplings are then put back in the consensus gauge to perform the comparisons described in the next sections using Eq. 7. The consensus gauge is the best for comparison because the statistics of the consensus symbol are the easiest ones to measure accurately.

B. Removing interactions

An alternative, complementary regularization scheme consists in reducing the number of interactions to be inferred from the data. Sparsification of the interaction network is sometimes achieved through L1 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. 3 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, 31], 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

$$\epsilon_{\rm p} = \frac{1}{Nq}\sqrt{\sum_{i,a}\left(\frac{\Delta f_i(a)}{\sigma_i(a)}\right)^2}. \tag{12}$$

t	ϵ_{max}	K_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: Local minima of the maximal relative error for the reference model: ER05, of Fig. 3 data sample of B=1000 configurations, and cut frequency $f_0 = 0.01$. The Table gives the cluster inclusion threshold t, the number K_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. The entropy of the model having generated the data is $S = 50.3$. The optimal threshold determined by the spACE procedure is 3.6×10^{-2}.

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

$$\epsilon_{\rm c} = \frac{2}{N(N-1)q^2}\sqrt{\sum_{i<j,a,b}\left(\frac{\Delta c_{i,j}(a,b)}{\sigma^c_{i,j}(a,b)}\right)^2}, \tag{13}$$

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

$$\epsilon_{\rm max} = \max_{\{i,j,a,b\}} \frac{1}{\sqrt{2\log(M)}}\left(\frac{|\Delta f_i(a)|}{\sigma_i(a)}, \frac{|\Delta f_{ij}(a,b)|}{\sigma_{ij}(ab)}\right), \tag{14}$$

where $M = Nq + (N(N-1)/2)q^2$ is the total number of one- and two-point correlations. As shown in Fig. 3 (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. 3.

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. 3: 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. 3 is given by the number K_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 $K_2 = 55$ 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 K_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. 3). 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 $K_2 = N \times (N-1)/2$, 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 $K2_m$ of 2-site clusters and records the inferred parameters at the relative minima of the maximal error. This procedures imposes a prior knowledge on the sparsity of the interaction graph by giving an upper

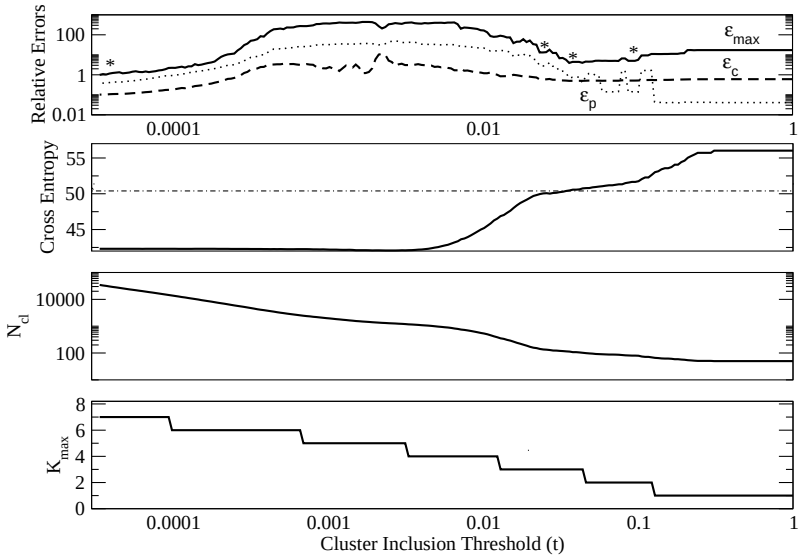


FIG. 3: Cluster expansion as function of the cluster inclusion threshold t for one data set, obtained by sampling one realization of an Erdős-Rényi random graph with $B = 1000$ configurations and applying ACE on the color compressed data with $f_0 = 0.01$. From top to bottom: *i)* relative reconstruction errors versus t on frequencies ϵ_p, connected correlations ϵ_c and maximal relative error ϵ_m. Stars indicate possible solutions of the inverse models corresponding to $\epsilon_m \approx 5$ *ii)*, Regularized cross entropy versus t, *iii)* number of total clusters included in the expansion versus t, *iv)* maximal cluster size versus t.

bound for the number of edges. As an illustration, 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 $K2_m = N\ 2.5 \approx 125$. The spACE procedure is robust with respect to the $K2_m$ value used. For the data and the model of Fig. 3, any value of $K2_m < N\,(N-1)/2$ rules out the fully connected minimum and allows one to select the best sparse model obtained at the threshold $t = 0.036$ of Table I. In the following we have therefore stopped the algorithm for $K2_m = 100$ and $K2_m = 200$, and we have verified that results are stable for the two values. his procedure greatly reduces the computational time, which increases linearly with the number of computed clusters and grows exponentially with their size as q^{K2_m} (see Sec. IV E and Appendix VII B).

In practice, to find the best sparse graph, with a number of edges smaller than the prescribed value $K2_{max}$, we adopt the following procedure. For each interval in threshold values corresponding to a threshold decreasing of a factor $\tau = 3.4$ [37]. The model parameters giving the minimal relative error are recorded, and the algorithm is stopped when the number of 2-site clusters summed up in the expansion is equal to $K2_{max}$. Recording the minima and the number of 2-site clusters $K2$ the different threshold intervals allows one to track the sparsity of the inferred graphs better reproduce the data. Among the recorded model the set of parameters and the inferred graph giving the minimal errors are chosen.

IV. BENCHMARKING ON SYNTHETIC DATA

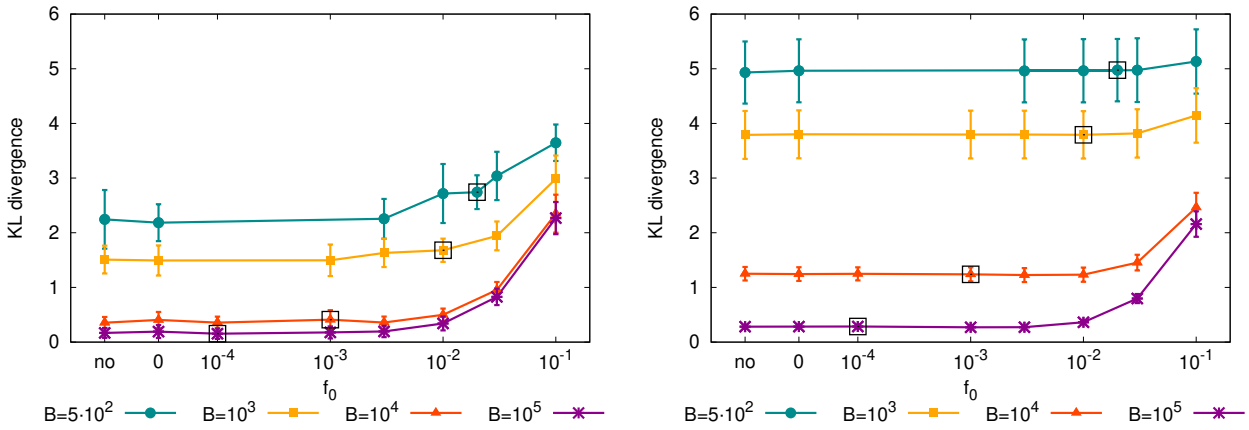
To carry out an extensive analysis of the effects of the color compression introduced in Section III A on the quality of the inference, we will apply it to artificial data generated by Potts models on Erdős-Rényi (ER) random graphs. The model and the generation of the data are described in Section III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

Once the data are obtained, we apply the compression schemes introduced in Section III A 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



(a) ACE (b) PLM

FIG. 4: 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 a cutoff frequency $f_0 = 10/B$.

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) [38], 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 VII 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. 3 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

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}.

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. 3 and in table I obtained both with importance sampling and the ACE expansion.

The fully connected graph has a larger KL divergence and it is therefore overfitting the data while the sparse graph better reproduces the original model. All results for the ACE expansion presented in the following are obtained by 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.

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

Fig. 4 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 ER models, which are indeed sparse by construction. It is worth noticing that the KL divergence for the fully connected model inferred by PLM for the reference case ($B = 1000$, $f_0 = 0.01$) is larger than the one obtained for the sparse ACE model, but smaller than the one for the fully connected ACE graph of Table II: PLM inference with large regularization gives better model reconstruction than with small regularization, see Table IV in Appendix VII D. The KL divergence is very stable for low frequency thresholds, starting to grow only at frequency thresholds $f_0 \sim 0.01$. This increase is of course much more significant at high B; indeed, reducing the number of explicitly modeled Potts states results in a loss of information affecting the quality of the inference, when the sampling is good. The black empty squares correspond to cut frequencies $f_0 = 10/B$, or equivalently grouping symbols observed less than 10 times. This seems to be the frequency cutoff above which the performance of the inferred model is poor.

The results shown above for PLM were found using strong regularization ($\gamma_J \sim N/B$); Results for low regularization can be found in Appendix VII D.

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 $f_i(a) = \sum_{generated\ \mathbf{a}} \delta(a_i, a)/B_{gen}$

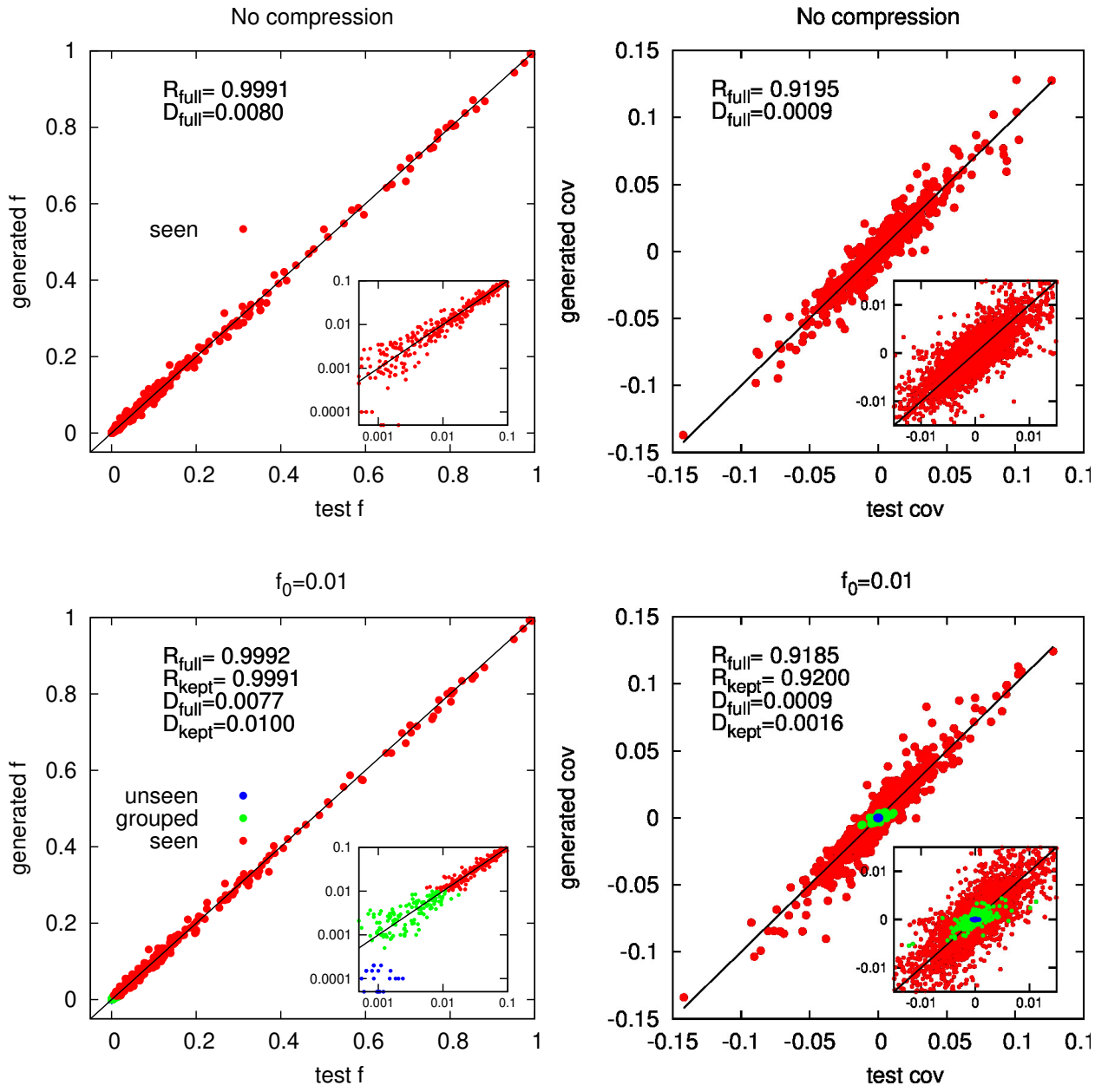


FIG. 5: Reconstruction of average frequencies and covariances. Comparison between generated data and test set for no color compression (top) and $f_0 = -0.01$ for ACE. The Pearson correlation coefficient (R) and the absolute error (Δ, Eq. 16) are marked on top of the plots both for the full model and the for the reduced one (only explicitly modeled states).

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.

Figures 5 and 6 show, 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

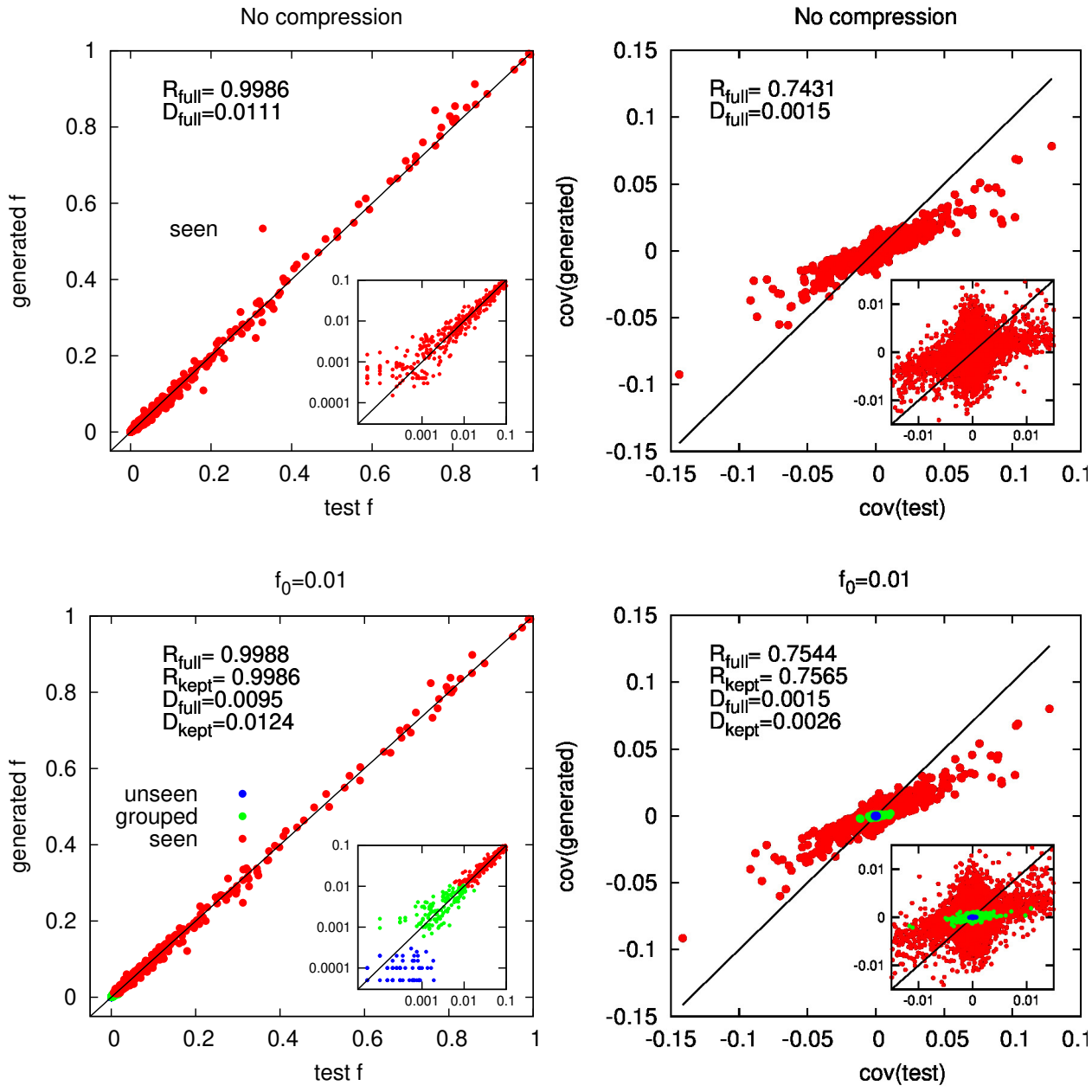


FIG. 6: Reconstruction of average frequencies and covariances. Comparison between generated data and test set for no color compression (top) and $f_0 = -0.01$ for PLM. The Pearson correlation coefficient (R) and the absolute error (Δ, Eq. 16) are marked on top of the plots both for the full model and the for the reduced one (only explicitly modeled states).

panels). As is shown in the figures the PLM covariances are dowscaled due to the strong regularization, as happens for the couplings (Section IV D). Moreover PLM assign smaller frequencies to the unseen Potts states (left panels of Fig. 6 in log-log scale) this is probably due to the fact that the pseudocount used during decompression seems to be well fixed for the Bayesian regularization used in ACE but not for the large regularization used in PLM. The inset in Fig 6 shows that, contrary to spACE (inset of Fig. 5), zero covariances are set to non-zero values with PLM because of overfitting.

To have a more systematic comparison, we analyzed the Pearson correlation coefficient R of frequencies

and covariances, as well as their absolute error defined as:

$$\Delta f = \sqrt{\frac{\sum_i \sum_a \left(f_i^{gen}(a) - f_i^{test}(a)\right)^2}{\sum_i q_i}}\,, \tag{15}$$

$$\Delta cov = \sqrt{\frac{\sum_{ij} \sum_{ab} (cov_{ij}^{test}(a,b) - cov_{ij}^{gen}(a,b))^2}{\sum_{ij} q_i \cdot q_j}}\,. \tag{16}$$

These quantities are then computed for different B and f_0 and averaged over 10 graph realizations. The results are shown in Figs. 7 and 8.

Performance seems again stable at low f_0 before progressively degrading at large f_0 even if large sample to sample fluctuations are present for smallest sampling depths. spACE outperforms PLM for covariance reconstruction, especially at small sample depth.

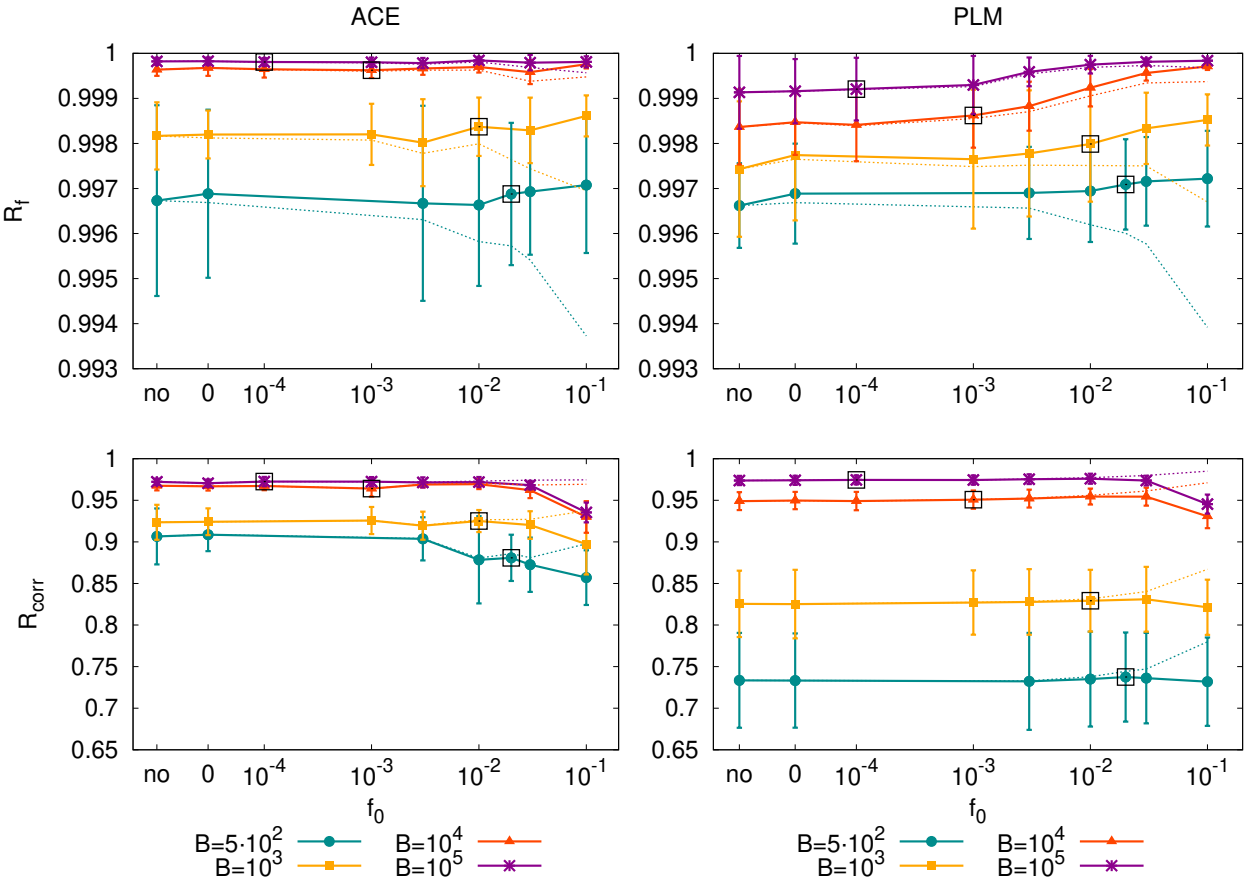


FIG. 7: Pearson correlation between test and generated frequencies (R_f, top panels) and covariances, (R_{cov}, bottom panels) averaged over 10 ER realizations as a function of the color compression for several sample sizes. Dashed lines: Pearson correlations restricted to the explicitly modeled Potts states. Full lines: Pearson correlations on all states. Error-bars are standard deviations computed over the 10 realizations. Inference is performed respectively by ACE (left) and PLM (right).

C. Interaction networks

In this section we focus on the reconstruction of the interaction network and the prediction of pairs of sites that are interacting, or "in contact", in the interaction network. The original ER graph is sparse, with

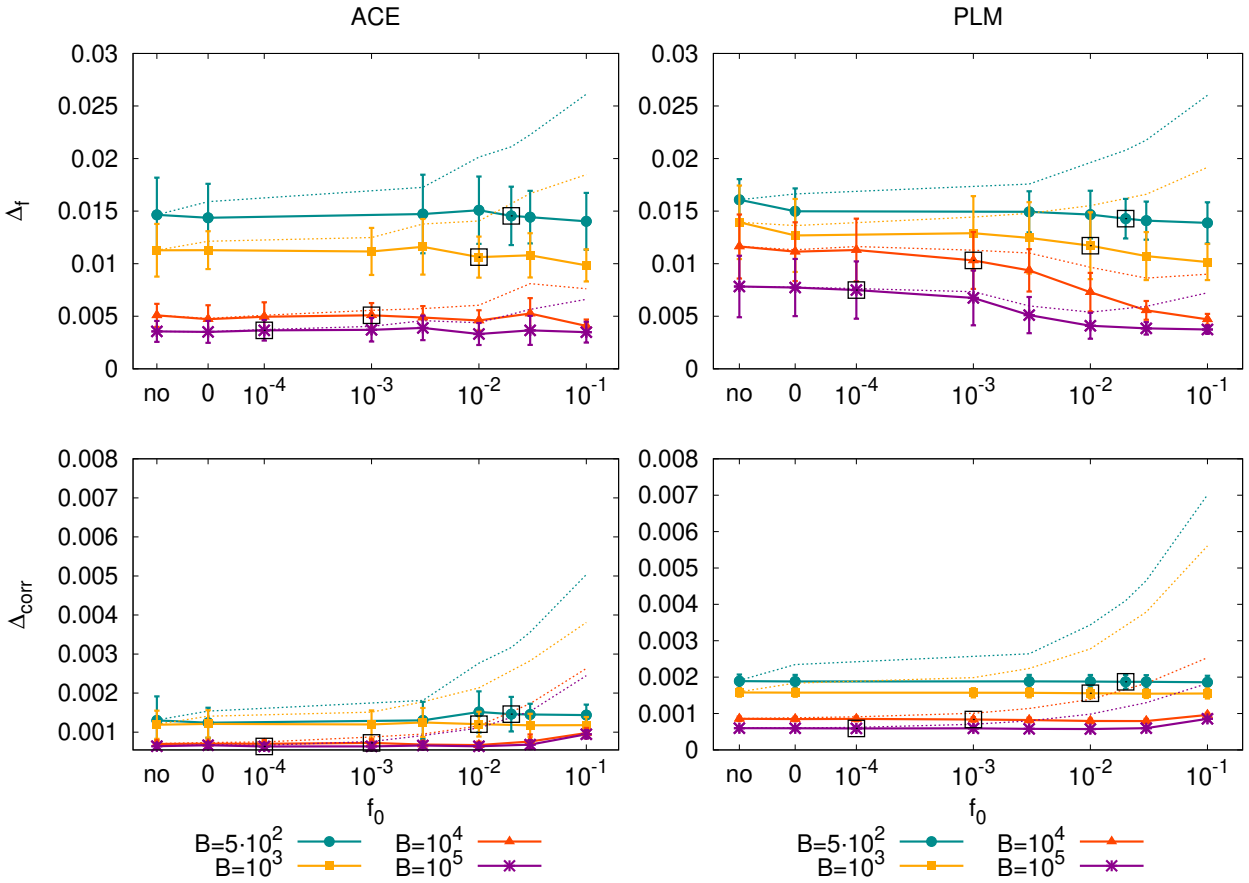


FIG. 8: Absolute error (Eq. 16) of frequencies (Δ_f, top panels) and covariances (Δ_{cov}, bottom panels) averaged over 10 ER realizations, as a function of the color compression for several sample sizes. Dashed lines: error on explicitly modeled Potts states only. Full lines: error on all parameters. Error-bars are standard deviations computed over the 10 realizations. Inference is performed respectively by ACE (left) and PLM (right).

an average connectivity of about 2.5 (see Fig. 1). We can predict contacting sites as those site pairs with large couplings, as traditionally done for protein structures [15–17]. To this end, we compute the Frobenius norm of the (10×10) inferred and decompressed coupling matrix between each pair of sites i, j,

$$F_{ij} = \sqrt{\sum_{a,b} J_{ij}(a,b)^2}\,. \tag{17}$$

We first show the results with ACE and PLM for the reference model, ER05 with configuration sampling $B = 1000$ and compare a single frequency cut $f_0 = 0.01 = 10/B$ to the decompressed case. In the second part of the section we compare the results for several sample sizes and frequency thresholds by averaging over multiple graph realizations.

In Figs. 9 and 10 (top panels) we compare the real network with the Frobenius norms of the couplings inferred by ACE and PLM with no color compression (left) and $f_0 = 0.01$ (right). With ACE, inferred with the spACE procedure, the inferred network is sparse with only a limited number of sites being adjacent. In the example of Fig. 9, without color compression (top left panel), $N_{pred} = 49$ sites have a nonzero Frobenius norm over N_0=59 sites linked by edges in the original graph (see Fig. 1). 16 edges are missed, while 6 site pairs are falsely predicted to be in interaction. Instead, for $f_0 = 0.01 = 10/B$ (top right panel of Fig. 9) only 14 adges are missed but 10 site pairs are wrongly predicted to be in interaction, implying a slight degradation

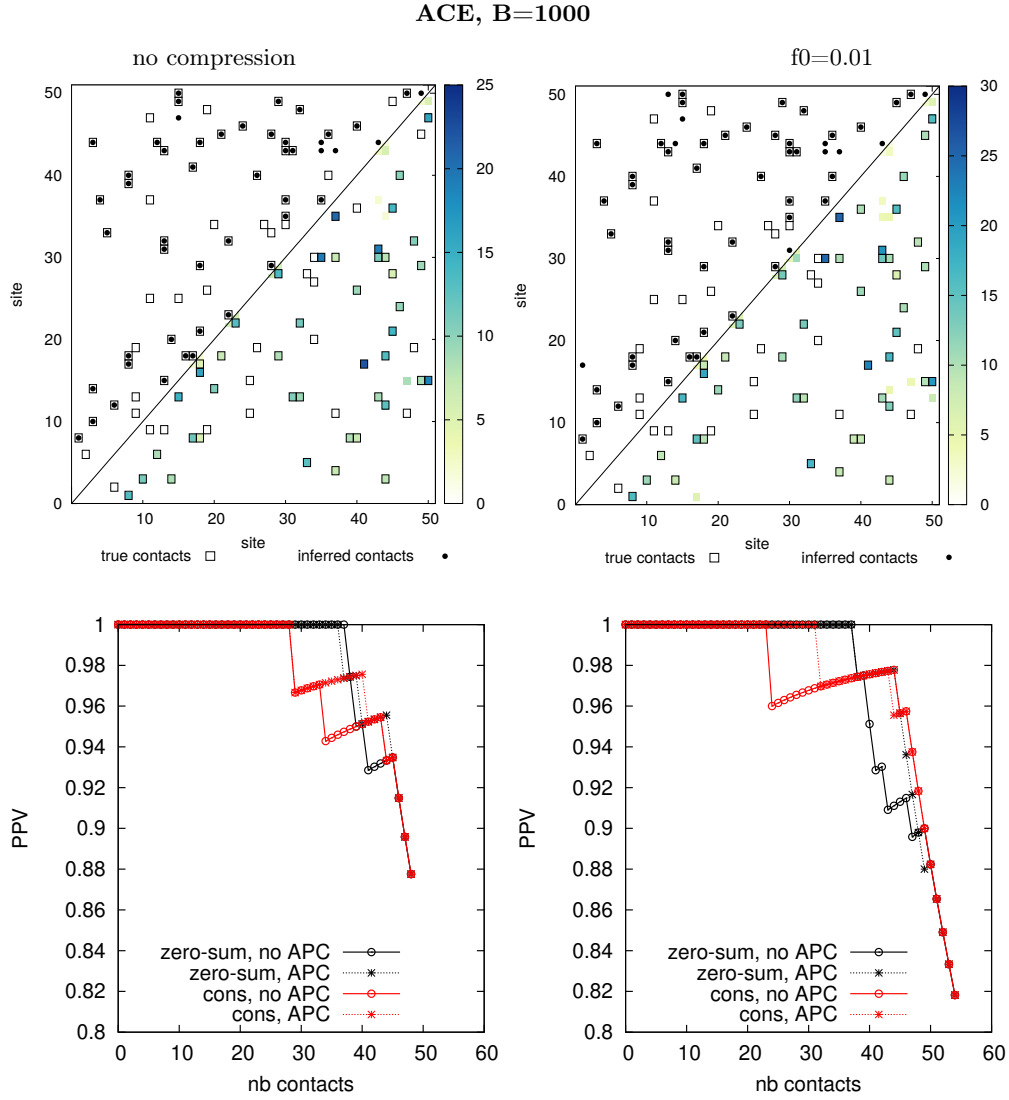


FIG. 9: ACE contact map reconstruction and PPV curve for one realization of ER graph. Left: no color compression. Right: $f_0 = 0.01$.
Top: contact maps. Upper triangular: contact map with real contacts (black empty squares), inferred true positive (full circles in empty squares), inferred false positive (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.
Bottom: Positive Predicted Value (PPV) curve in consensus and zero-sum gauge with and without APC.

of the precision but an improvement of the recall.

Contrary to ACE, PLM with the L_2 norm regularization described in Eq. 5 infers a fully connected interaction graph: all the $N(N-1)/2$ Frobenius norms are different from zero as shown in Fig.10 (bottom panel). There is therefore no straightforward separation between pairs of sites predicted to be in interaction or not. In the top panels of Fig. 10 we put a dot for predicted interactions, hereafter referred to as contacts, only for the N_0 site pairs with highest Frobenius norm, where N_0 is the true number of edges. In the example shown in the figure, there are 42 true positive, 17 false positive, and 17 missed edges both without color compression and with $f_0 = 0.01$. These results are comparable to the ones obtained by ACE in Fig.9.

To gain more insight into these predictions, as done for protein structure [25, 28, 39], we can sort site pairs by decreasing Frobenius norm and follow the precision obtained progressively including the corresponding

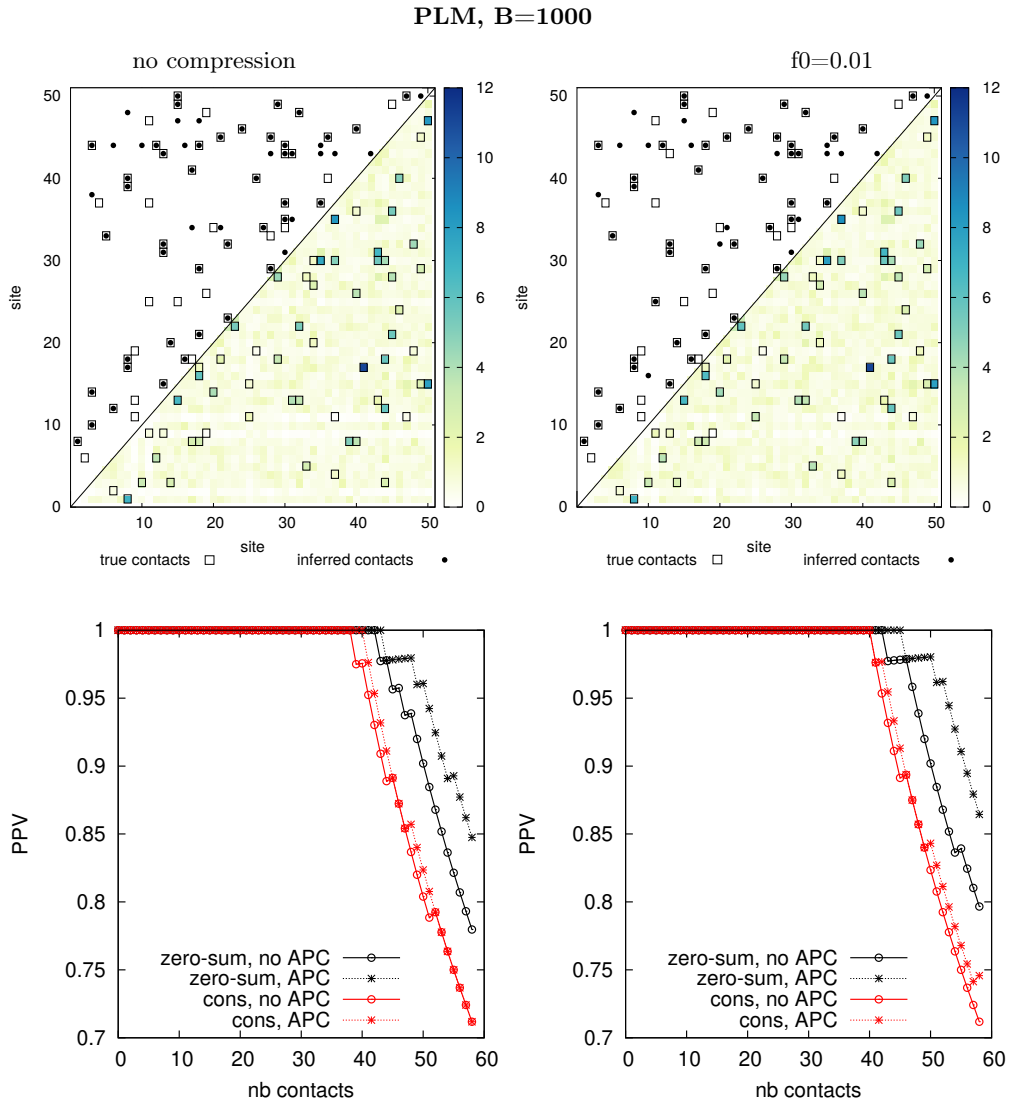


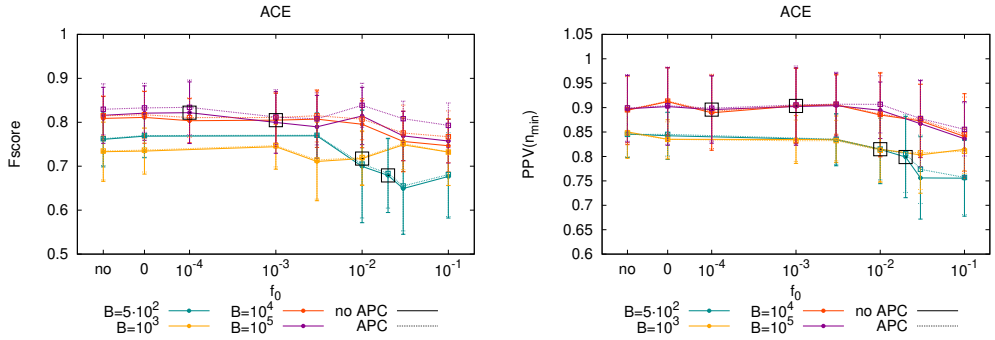
FIG. 10: PLM network reconstruction and PPV curve for one realization of ER graph. The site pairs with highest Frobenius norm up to the number of real edges in the graphN_0 are considered predicted contacts. Left: no color compression. Right: $f_0 = 0.01$.
Top: contact maps. Upper triangular: contact 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.
Bottom: Positive Predicted Value (PPV) curve in consensus and zero-sum gauge with and without APC.

site pairs in the so called Positive Predicted Value (PPV) curve:

$$PPV(n) = \frac{TP(n)}{n}. \tag{18}$$

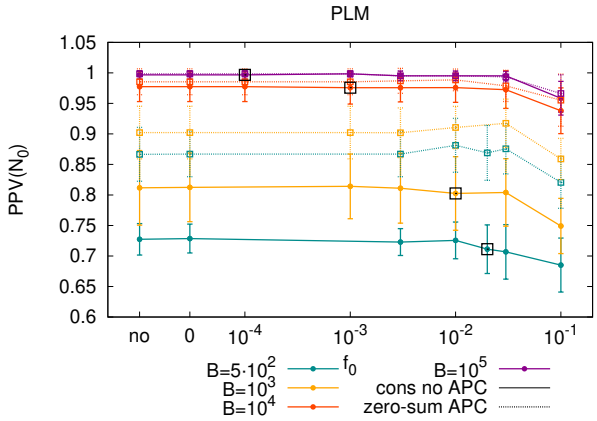
Where $TP(n)$ is the number of true predicted edges in the top n pairs. This is shown in the bottom panels of Figs. 9 and 10 for site pairs up to the last with non-zero norm (for ACE) or to N_0 (for PLM).

In the top panels the Frobenius norm is computed in the consensus gauge, i.e. gauging to zero the most frequent state on each site in the considered configuration sample, because this is the gauge in which we will then compare couplings and fields in the next section. However, it has been empirically shown on protein



(a) ACE's Fscore

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



(c) PLM $PPV(n_{min})$

FIG. 11: Fscore and $PPV(n = min\{N_0, N_{pred}\})$ 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).

structure prediction that inferred edges are more precise in the zero sum gauge and that the performance can be further improved with the Average Product Correction (APC) [20, 25, 41],

$$F_{ij}^{APC} = F_{ij} - \frac{F_{i.}\, F_{.j}}{F_{..}}\,. \tag{19}$$

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. The APC decreases the norm of those pairs where at least one site has large norms with many others, this being possibly due to undersampling. The comparison of the PPV in Figs. 9 and 10 is then done in both gauges, with and without the APC. Note that, with ACE, 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.

We must average over different realizations in order to have a more significant comparison and to distinguish statistical fluctuations from systematic worsening of the performance. For PLM, due to the lack of an explicit separation in the Frobenius norms, we adopt the positive predicted value (PPV) at the number of real contacts $PPV(n = N_0)$ as the quality measure. For ACE, where a clear separation between predicted contacts and predicted non-contacts is possible, we use two quality measures:

- To encompass both the precision and the recall in a single measure we use the Fscore, which is the

harmonic mean between the two and gives

$$\text{Fscore} = 2\frac{TP(N_{pred})}{N_{pred} + N_0} \tag{20}$$

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;

- We also compute PPV at the lesser of the number of predictions N_{pred} or real edges N_0.

The quantities above are shown in Figs. 11a, 11b and 11c respectively for ACE and PLM as a function of the color compression. 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 are due to the sparsity threshold. We have verified that, at such sampling for fully connected network, ACE has the same PPV as PLM: for the reference graph sampled with $B = 10^4$ and $p_0 = 0.01$, we obtain $PPV(N_0) = 0.93$, and, using APC correction, $PPV_{APC}(N_0) = 0.95$ at low threshold and $PPV(N_0) = 0.87$ ($PPV_{APC}(N_0) = 0.88$) at sparse threshold. We see that, for both algorithms, the performance is usually stable against the introduction of color compression.

D. Couplings and Fields

In Figs. 12 and 13 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 (Fig. 12) and PLM (Fig. 13). Different colors in Fig. (Fig. 12) and (Fig. 13) 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.

Fig. 12 shows that, as observed in Section IV C the sparse procedure, spACE, misses some edges and the corresponding couplings are fixed to zero. As shown in Fig. 13, 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.

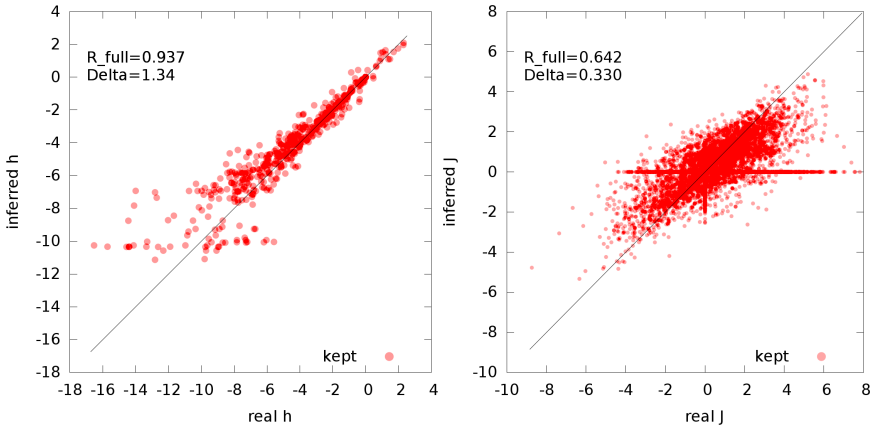
To analyze the results in a more quantitative way, we compute the Pearson correlation coefficient R between the real and the inferred parameters ($R_* = \frac{cov(real\,*, inferred\,*)}{\sigma_{real\,*} \cdot \sigma_{inferred\,*}}$ for $* = h, J$) and the couplings and field absolute errors, defined as:

$$\begin{aligned} \Delta h &= \sqrt{\frac{\sum_i \sum_a \left(h_i^{inf}(a) - h_i^{real}(a)\right)^2}{\sum_i q_i}}\,, \\ \Delta J &= \sqrt{\frac{\sum_{ij} \sum_{ab} (J_{ij}^{real}(a,b) - J_{ij}^{inf}(a,b))^2}{\sum_{ij} q_i \cdot q_j}}\,, \end{aligned} \tag{21}$$

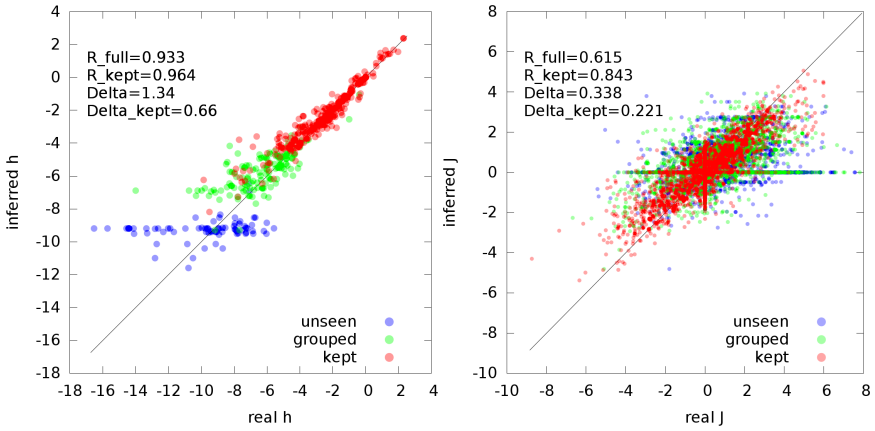
which measures the average distances from the diagonal of the points in the scatter plots of Fig. 12 and Fig. 13. Figs. 14 and 15 show the Pearson correlation coefficients and the absolute errors for various B and as a function of f_0, averaged over 10 ER realizations. Here, the full line indicates the behavior for the full decompressed model, while the dotted line indicates performances only for states explicitly modeled during inference. From both figures it is evident that ACE gives better results than PLM for parameter reconstruction, especially on couplings. 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.

Fig. 14 and 15 show that the performances are stable as a function of the color compression up to a value of f_0, where performances drop because the compression become too strong.

The dashed lines in Figs. 14 and 15 show that by restricting the coupling comparison to better and better sampled states, at large f_0, the reconstruction indicators are better and better. On the contrary for the reconstruction of fields performances are better when comparing all the Potts states, because correctly



(a) ACE: no color compression



(b) ACE: $f_0 = 0.01 = 10/B$

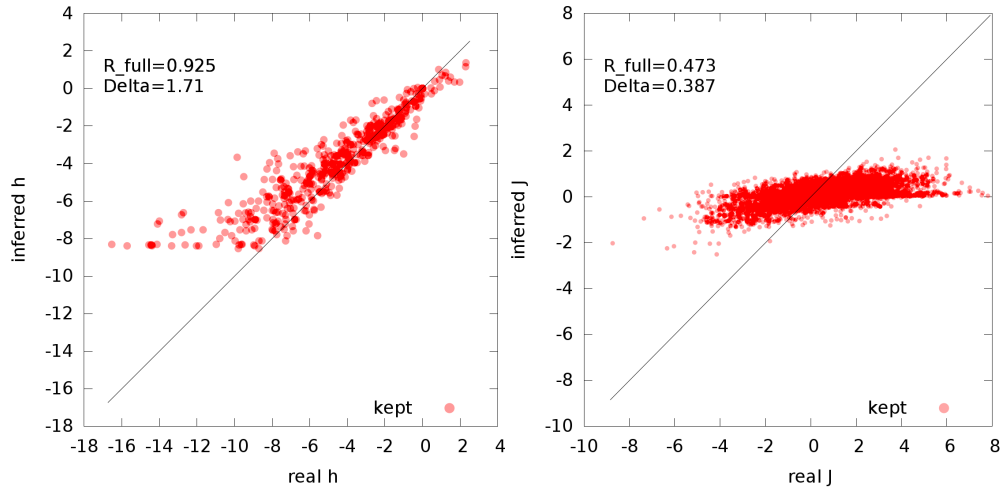
FIG. 12: Comparison of inferred and real fields and couplings with ACE for one realization of ER graph with no color compression (top) and $f_0 = 0.01$ (bottom), for B=1000 sampled configurations. Parameters on explicitly modeled (kept), grouped, and unseen Potts states are colored differently. Left: field comparison. Right: coupling comparison. On top of each plot, the Pearson correlation coefficient (R) and the absolute error (Δ, as in Eq. 21) are indicated.

reconstruct, through the decompression procedure of Sec.III A the large and negative fields for the grouped and unseen Potts symbols. The overall reconstruction of couplings and fields is both for PLM and ACE stable below cutoff frequencies $\frac{f_0 \simeq 10}{B.}$

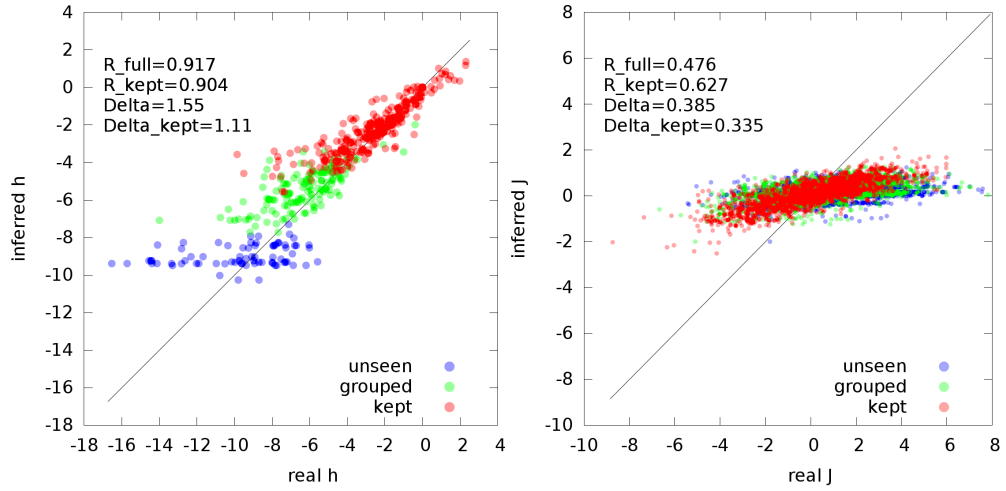
E. Gain in computational time

Figure 16 shows how the computational time scales with the sample size (top) and the color compression frequency threshold (bottom) for ACE (left panels) and PLM (right panels). These times have been obtained on a processor Intel® Xeon(R) CPU E5-2690 v4 @ 2.60GHz x 56 and are shown for for a single ER graph realization. The computational time for standard ACE depends on three factors: q^K where K is the cluster size and q the number of Potts states, the overall number of clusters in the construction rule, and the number of Monte-Carlo steps to calculate the relative errors in the reconstruction of the first and second moment of the data distribution.

The effect of compression on the runtime therefore depends on the choice of stopping conditions and Monte-Carlo steps when running the algorithm. At small cluster sizes, Monte-Carlo sampling is the dominant



(a) PLM: no color compression



(b) PLM: $f_0 = 0.01 = 10/B$

FIG. 13: Comparison of inferred and real fields and couplings with 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. 21) are indicated.

contribution to the runtime, which does not depend strongly on the number of colors. At large cluster sizes, computing the partition function and numerically maximizing the likelihood become the dominant contributions to the runtime, in which case color compression can provide substantial benefits. For example, inference of models for many HIV proteins [42–45] would take prohibitively long times without significant compression. In the spACE implementation used for this paper the limiting step is the large number of Monte Carlo steps (500000) used to calculate the relative errors with high precision , which does not depend on the number of parameters, thus diminishing any time reduction with color compression. Having a large number of MC steps is in important to correctly estimate the relative errors at very large sample size, given the small value of the sampling variances.

For PLM (right panel of Fig. 16), time increases linearly with the number of parameters to infer and almost linear in the sample size B. Time can then be reduced thanks to color compression. So, while spACE is optimal for large sample sizes, PLM is faster for small B, especially when color compression is used.

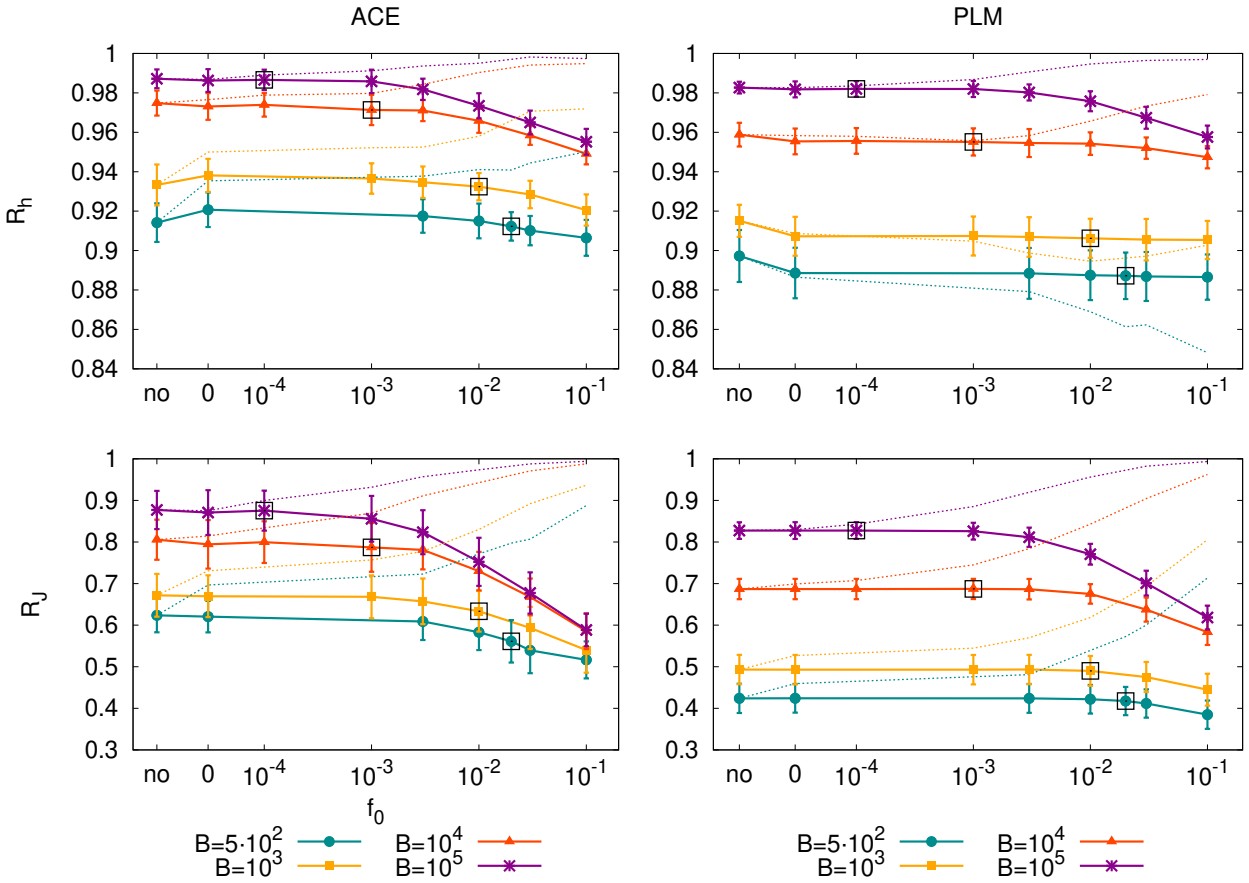


FIG. 14: Pearson correlation between real and inferred fields, (R_h, top panels) and real and inferred couplings, (R_J, bottom panels) averaged over 10 ER realizations as a function of the color compression for several sample sizes. Dashed lines: correlations between inferred and real parameters restricted to the explicitly modeled Potts states. Full lines: correlations on all parameters, after decompression of unseen and grouped Potts states (see Sec. III A). Error-bars are standard deviations computed over the 10 realizations. Inference is performed respectively by ACE (left) and PLM (right).

V. COLOR COMPRESSION APPLIED TO SEQUENCE ALIGNMENTS 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 [16, 39] or the fitness landscape [21–24]. In particular, we would like to compare the change sof energy corresponding to single point mutations with respect to a wild-type protein sequence to the experimentally measured changes of fitness of the protein.

We here consider three protein families whose fitness has been systematically assessed against single-point mutations:

- WW 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 is a protein domain present in signaling proteins. Here fitness has been measured in terms of binding affinity [47];

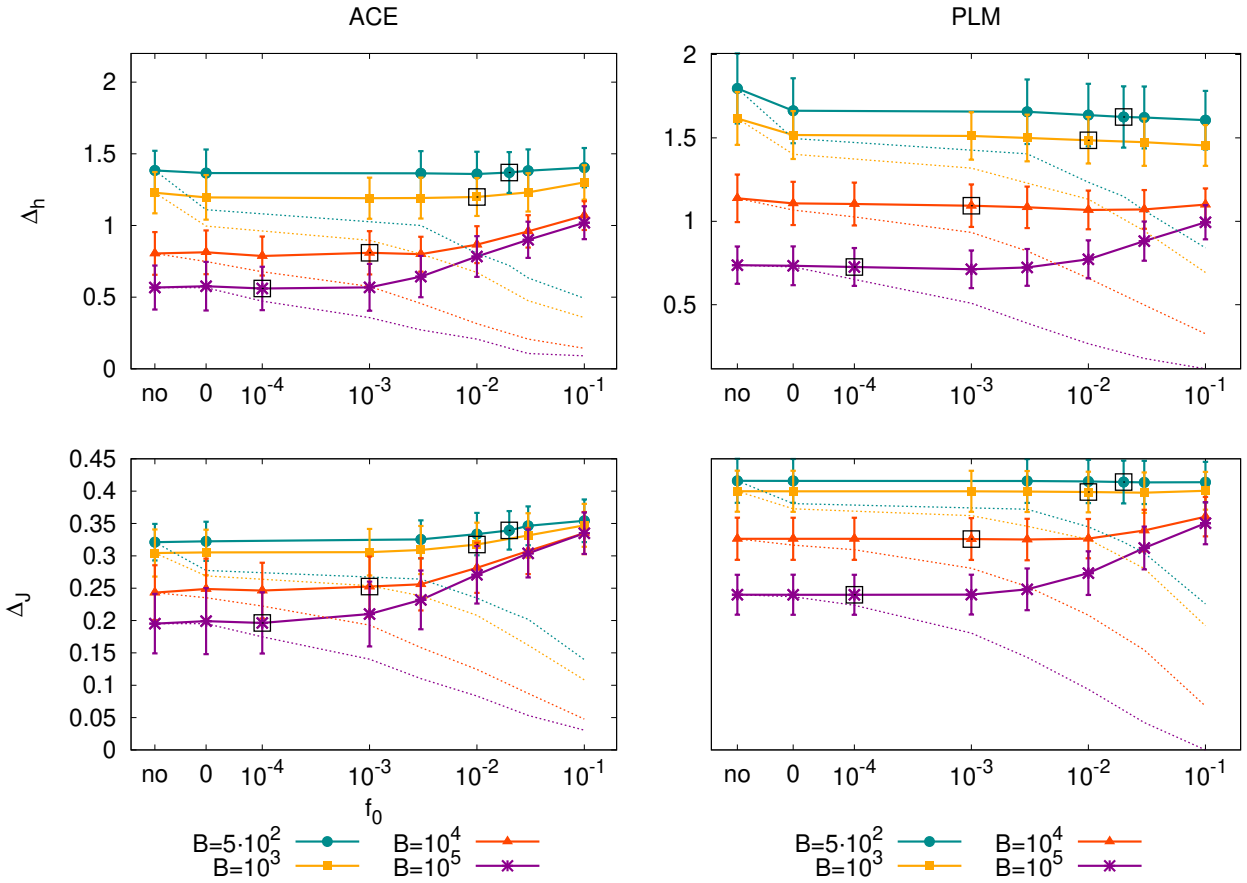


FIG. 15: Absolute errors (Eq. 21) on fields (Δ_h, top panels) and couplings (Δ_J, bottom panels) averaged over 10 ER realizations, as a function of the color compression for several sample sizes. Dashed lines: error on parameters related to explicitly modeled Potts states. Full lines: error on all parameters, after decompression of unseen and grouped Potts states (see Sec. III A). Error-bars are standard deviations computed over the 10 realizations. Inference is performed respectively by ACE (left) and PLM (right).

- RRM is an RNA recognition motif; fitness was estimated through growth rate measurements in [48].

Alignments and experimental fitness measures used in this section have been retrieved from a recent paper [24].

Figure 17 summarizes the impact of color compression on inference for these three cases. 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 top right corner of fig. 17 shows the variation of the Spearman correlation coefficient as a function of the color compression. The SpACE procedure was here applied with $K_{2m} = 2N$; however the relative error ϵ_{max} (Eq.14) was generally too large even at its local minima, indicating that the procedure has not converged. As shown in [31] a Boltzmann Machine Learning (BML) procedure was further used, starting from the spACE inferred parameters as initial guess, to better reproduce the low order statistics of the data and therefore the quality of the inferred model. ACE + BML performances are overall compatible with PLM results, and are better for well sampled case such as RRM, and slightly worse in the PDZ case. A certain level of color compression does not globally harm the performances, neither for PLM nor for ACE.

The top left panel in Fig. 17 shows the reduction in number of parameters due to color compression and the bottom panels show the reduction in computational time for ACE and PLM respectively. We observe that this reduction is much stronger than for the ER case analyzed above, mainly due to the much larger

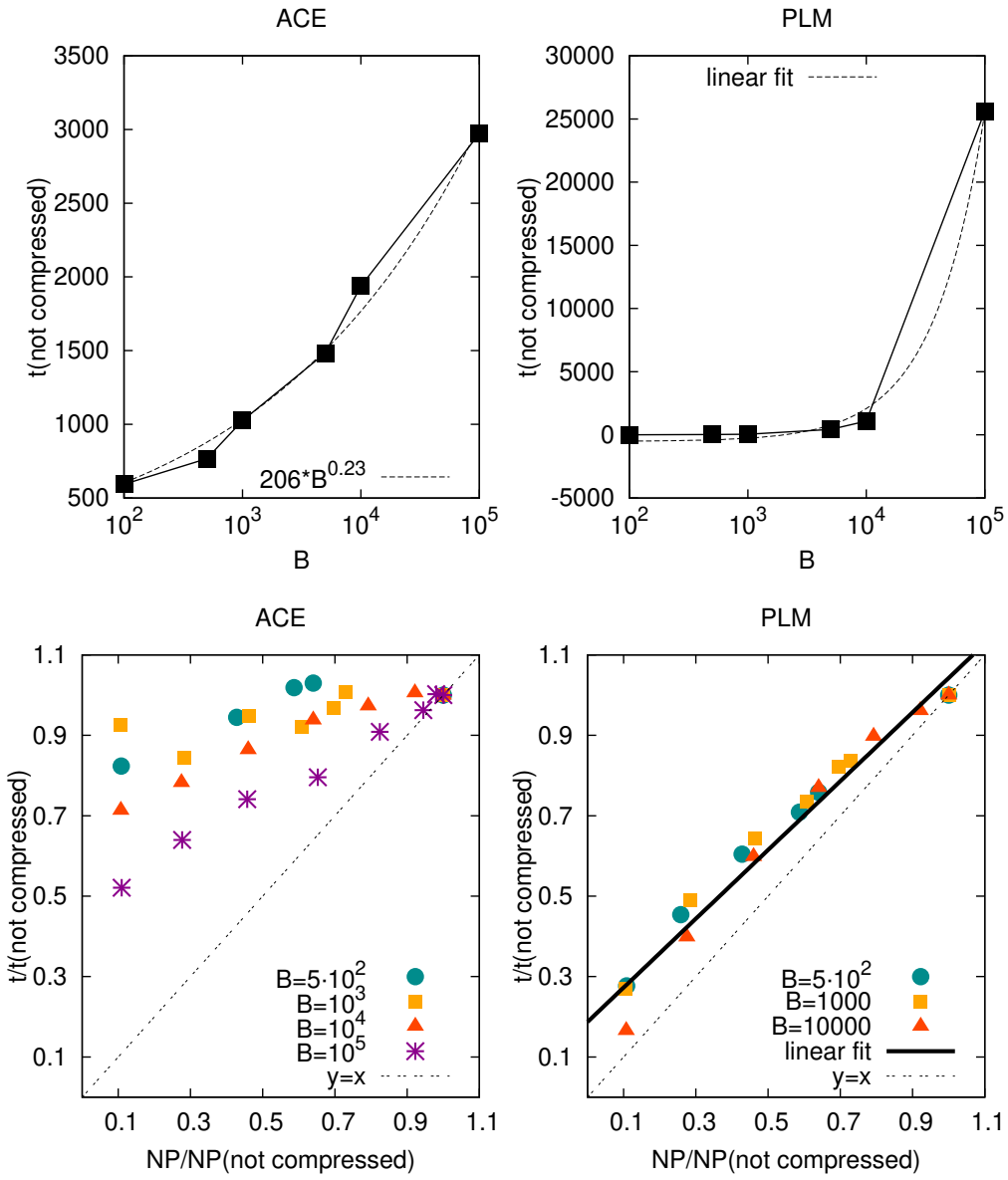


FIG. 16: Computational times for ACE and PLM for different B and color compression. Top: computational time, in seconds, using only 1 CPU, for ACE (left) and PLM (right) as a function of the sample sizes B; a power law and a linear fits are added respectively to ACE and PLM (dashed line). Bottom: computational time ratio between the compressed and decompressed inference for ACE (1 CPU; left panel) and PLM (25 parallel CPU; right panel) as a function of fraction of parameters to infer; on top of it the diagonal is drawn in dotted line and, for PLM, the linear fit (full line) is also added.

number of possible states ($q = 21$) of proteins. We even observed that ACE does not always converge for real proteins without color compression, because the algorithm gets stuck in trying to recover the statistics on the rarest states, while this almost never happens when the rarer states are grouped. The significance of this reduction is even clearer when looking at absolute running times in Table III.

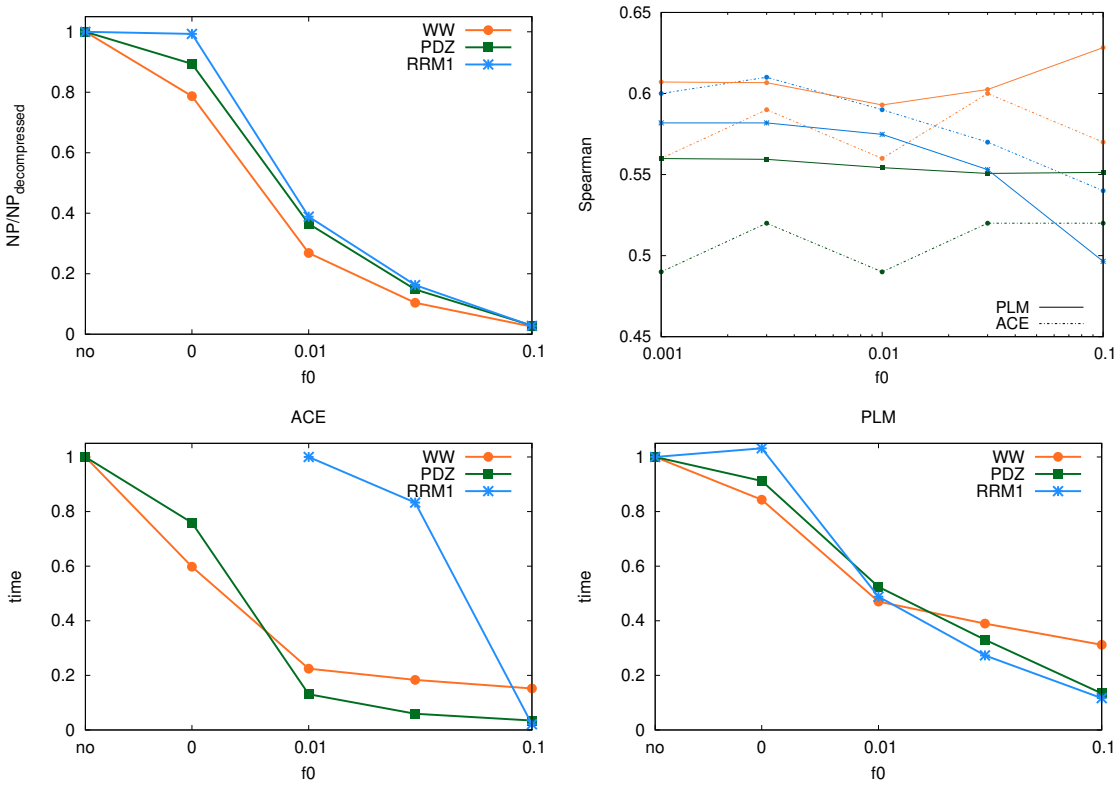


FIG. 17: Color compression on real proteins with ACE and PLM. The top panels show respectively the fraction of explicitly modeled parameters as a function of f_0 (top left) and the Spearman correlation coefficient between the fitness predictions and the experimental measures found in literature, as a function of the color compression (top right). The bottom panels show the time gain due to color compression in these same runs respectively for ACE (bottom left) and PLM (bottom right). The protein families used here are: WW (PF00397, in orange), PDZ (PF00595, in green) and RRM1 (PF00076 in light blue).

protein	$f0$	t_{ACE}	t_{PLM}
WW	no	1668	220
WW	0	997	206
WW	0.01	375	85
WW	0.1	253	17

protein	$f0$	t_{ACE}	t_{PLM}
PDZ	no	38303	649
PDZ	0	29076	592
PDZ	0.01	5009	340
PDZ	0.1	1324	87

protein	$f0$	t_{ACE}	t_{PLM}
RRM1	no		2927
RRM1	0		3020
RRM1	0.01	64938	1429
RRM1	0.1	1350	341

TABLE III: Example of running times for ACE and PLM (in minutes on a Desktop computer) for the three studied protein families for different color compression thresholds f_0.

VI. CONCLUSION

We have benchmarked the inference of color-compressed Potts models from data generated by Potts models on Erdős-Rényi random graphs. In such compressed inference, the poorly sampled colors are lumped together in a unique and effective Potts state and therefore the number of Potts states explicitly modeled depends on the site. Knowing the 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 have focused on the undersampling regime, where the number of the parameters ($\sim 10^5$) is larger than or

equal to the number of data.

After the inference of the compressed model we have introduced a procedure to recover *a posteriori* a full Potts model. During the decompression procedure, the couplings of grouped or unobserved Potts variables are fixed to a reference value, while the values of the fields are adjusted according to the individual frequencies of the grouped symbols, or to a small pseudo-count for never observed Potts symbols. Such decompression has to to be carefully carried to stick to the chosen gauge as explained in Sec. III A. It is essential to ensure that the frequencies of non-observed Potts states are correctly reproduced and lower than the ones of observed states. Decompression is useful to compare the inferred model to the ground truth (when available), to assess the performance of the inference for different color compression strengths, and more generally when the model is used to predict the behavior of poorly sampled variables on the original data set.

Color compression does not affect the accuracy of the inferred model below a value f_0^* of the compression frequency cut-off, while largely reducing the number of variables to be inferred. At frequencies above f_0^* the inferred model is degraded with respect to the full model because well sampled variables are grouped and their individual correlations are lost. The cutoff value f_0^* is not very sharp: as a rule of thumb, it ranges between $f_0^* = 1/B$, where is the number of sampled configurations, and $f_0^* = 10/B$ meaning that symbols observed less than 1 to 10 times can be "safely" grouped in the compressed model. It depends slightly on the inference and decompression procedure, on the quantity we are looking at and on the sampling size, the smaller cutoff being better for bad sampled data ($B = 5 \times 10^2 - 10^3$).

When the regularization is properly chosen to avoid overfitting, color compression does not improve the inference performance but simply acts as a further regularization. As shown in Section IV A 2 and Appendix IVthis is not the case for PLM with small regularization for which the model reconstruction shows an optimal value of the compression. For well regularized inference procedures, the parameters of well sampled states are accurately inferred, independently of the presence of the grouped states, while those of poorly sampled states are essentially fixed by the regularization imposed either during the inference or during the color decompression. 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, as clearly shown by the performance restricted to the explicitly modeled symbols alone of Figs. 12 and 13. This underlines the difference between sites and states in a Potts model. In the standard renormalization procedure [49] 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. Such robustness holds for the two algorithms studied here, the Adaptive Cluster Expansion (ACE) and Pseudo-Likelihood Maximization (PLM), but may be not true for algorithms based e.g. on the inversion of the correlation matrix for which the number of zero modes of the correlation matrix can have drastic consequences on the accuracy of the inversion. The ACE variant introduced here, called spACE, selects a sparse solution to the inverse problem by imposing a maximal cutoff for the number of 2-site clusters included, through an appropriate choice of the cluster inclusion threshold t. For the PLM algorithm, we have adapted the routine of the group of Aurell and collaborators [9, 28] by adding color compression as described in this paper.

We plan to improve the spACE expansions in several directions. One possibility is to change the value of $K2_{max}$ imposed in the analysis depending on the sampling conditions, in such a way to better adapt the number of inferred parameters to the sampling. The procedure can be also improved, especially in view of the applications to real proteins, by lowering the threshold t at fixed number of 2-site clusters, i.e. to take into account the largest-size cluster on the same 2 site support clusters. Finally introducing a way to stop the cluster expansion without using direct Monte-Carlo sampling but on other reconstruction and sparsity criteria [50] would avoid Monte Carlo sampling and consistently speed up the computational time. When doing inference on the ER graphs considered here, ACE is not appreciably faster with color compression because of the sparsity and small size of the considered graph. However, for denser or larger networks, e.g. in application to real proteins, the computational time is dominated by the computation of the cross-entropy contributions of large clusters. In those cases color compression considerably speeds up the inference process, often becoming essential for solving the inverse problem in reasonable times [31]. With PLM, on the other hand, a reduction in the number of parameters always implies smaller computational times. It would be interesting in a forthcoming work, to compare on a similar data set ACE and a PLM inference with, instead of L_2-regularization, the L_1 norm regularization that imposes a sparse network. Yet, PLM with L_1-regularization differs from spACE; spACE, as described here, imposes sparsity on the interaction graph, but enforces a L_2-norm on the parameters for the selected clusters, and in particular for sites predicted to be in contact. An L_1^2 norm as the one introduced in [51] could provide a combination of sparsity and L_2-norm

that characterizes the ACE algorithm and may be optimal for inferences on sparse networks as in the present case. Inferring a sparse network is not only worthwhile for sparse original models but it improves model reconstruction in the large undersampling regime as shown in Fig. 4 [40, 50]. Moreover the reduction of computational time obtained thanks to color compression and sparsity is necessary when dealing with larger number of sites, e.g. whole genome inference [27, 52].

Last of all, let us emphasize that the color compression/decompression procedure introduced here is not restricted to pairwise graphical models. It could be used in other machine learning approaches, such as restricted Botzmann machines, recently shown to be powerful to identify constitutive amino-acid motifs in protein sequences [51].

Acknowledgements. We thank Lorenzo Posani for useful discussions.

VII. 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: $\Gamma = \{i_1, \dots, i_k\}, k \leq N$, we can formally write the cross-entropy as the sum of cluster contributions:

$$S(\mathbf{J}|\mathbf{f}) = \sum_{\Gamma} \Delta S_{\Gamma}, \tag{22}$$

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

$$\Delta S_\Gamma = S_\Gamma - \sum_{\Gamma' \subset \Gamma} \Delta S_{\Gamma'} \,. \tag{23}$$

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 (22) yields the cross entropy for the entire system of N variables. As detailed in [2, 31], 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. (22) by fixing a cluster inclusion threshold t and summing up in Eq.(23) only cluster contribution with $|\Delta S_\Gamma| > t$.

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

Fig. 18 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 of f_0, see the two stars in Fig. 18. Moreover the threshold t corresponding to the sparse inferred graph is largely independent of the level of color compression.

Fig. 16 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. 19 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. 19 the expansion was stopped to maximal relative error of order 10 at small threshold t after 11 days while it

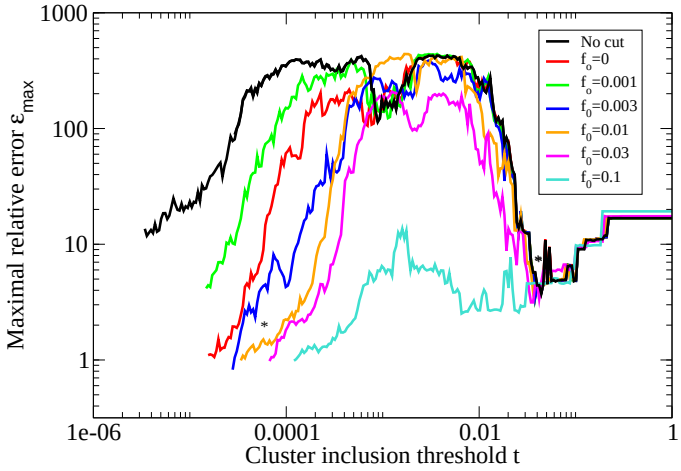


FIG. 18: Maximum relative error as a function of the expansion threshold for a particular graph realization (same used in Fig 5: ER05, sampled with B=1000),for different color compression f_0 .

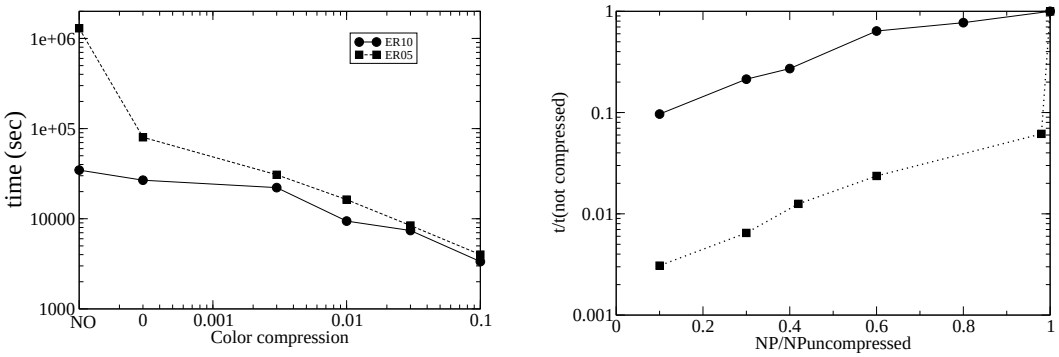


FIG. 19: Reduction in computational time due to the color compression for fully connected inferred models on 2 data sets obtained by sampling B=1000 configurations from two Erdős-Rényi random graph models. Left: Computational time at the low-threshold minimum as a function of the color compression threshold f_0. Right: Computational time relative to the one with no color compression as a function of the number of parameters.

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. 16 is smaller due to the reduced number of clusters. For the sparse graph, inference takes of the order of 16 minutes (on the same computer) independently of the color compression threshold, as shown in Fig. 16. For large interconnected models color compression can therefore be useful to reach convergence in a reasonable amount of time and infer a model that reproduces the statistics of the data.

C. 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

$$S_B = -\sum_{\boldsymbol{\sigma}} P_{\boldsymbol{J}^B}(\boldsymbol{\sigma}) \log P_{\boldsymbol{J}^B}(\boldsymbol{\sigma}) \ , \tag{24}$$

where the sum is over all possible configurations $\boldsymbol{\sigma} = \{\sigma_1, ..., \sigma_N\}$.

The inferred probability distribution at finite sampling B is

$$P_{\boldsymbol{J}^B}(\boldsymbol{\sigma}) = \frac{\exp\left(\sum_{i=1}^N h_i^B \sigma_i + \sum_{k<l}^N J_{kl}^B \sigma_k \sigma_l\right)}{\mathcal{Z}_B} \ . \tag{25}$$

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. (24) & (25) 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

$$\begin{aligned} S_{true} &= -\sum_{\boldsymbol{\sigma}} P_{\boldsymbol{J}^{true}}(\boldsymbol{\sigma}) \log P_{\boldsymbol{J}^{true}}(\boldsymbol{\sigma}) \\ &\approx S_{B\to\infty} = -\sum_{\boldsymbol{\sigma}} P_{\boldsymbol{J}^{B\to\infty}}(\boldsymbol{\sigma}) \log P_{\boldsymbol{J}^{B\to\infty}}(\boldsymbol{\sigma}) \ , \end{aligned} \tag{26}$$

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,

$$\begin{aligned} D(P_{\boldsymbol{J}^{true}}||P_{\boldsymbol{J}^B}) =& (S_B - S_\infty) + \sum_i h_i^B (\langle \sigma_i \rangle^B - \langle \sigma_i \rangle^\infty) \\ &+ \sum_{k<l} J_{kl}^B (\langle \sigma_k \sigma_l \rangle^B - \langle \sigma_k \sigma_l \rangle^\infty) \ , \end{aligned} \tag{27}$$

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 ($\gamma_J = \frac{1}{B}, \gamma_h = \frac{0.01}{B}$)
10^2	17.53	8.14	9.81
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 IV: KL divergences between inferred and empirical distributions for various regularization choices in ACE and PLM in the reference ER realization. No color compression applied.

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:

$$\begin{aligned} D(P_{\boldsymbol{J}^{true}}||P_{\boldsymbol{J}^B}) =& (S_B - S_\infty) + \sum_{i=1}^{N}\sum_{a=1}^{q} h_i^B(a)(\langle\sigma_{ia}\rangle^B - \langle\sigma_{ia}\rangle^\infty) \\ &+ \sum_{\substack{k,l=1\\k<l}}^{N}\sum_{c,d=1}^{q} J_{kl}^B(c,d)(\langle\sigma_{kc}\sigma_{ld}\rangle^B - \langle\sigma_{kc}\sigma_{ld}\rangle^\infty) \ . \end{aligned} \tag{28}$$

The artificial data are in a compressed representation (*cf.* Section III A). The complete inferred parameters are recovered as explained in Eq. (11).

D. KL divergence for PLM at low regularization and color compression

In this section we want to study what happens when using PLM at lower regularization, such as the $\gamma_J = 1/B$ used for ACE, as a function of the color compression. Without color compression, the performance obtained at $\gamma_J = 1/B$ becomes significantly worse, see Table IV. The main reason for this finding is that, at small regularization, one can obtain very large amplitude couplings between pairs of sites that are not actually interacting (true $J_{ij} = 0$ for all color pairs) due to low two-point frequencies $f_{ij}(a,b)$. These frequencies are affected by large statistical errors. If strong regularization can deal with them, we may hope that a similar improvement may be recovered at low regularization thanks to color compression. Indeed, by grouping together states that are not well sampled, the statistical errors on the two-point frequencies are reduced and this should lead to an improvement. At the same time we know that strong color compression lead to information loss on other well sampled two-point frequencies, compromising the performances. A more extensive discussion of the choice of the regularization related to protein sequence analysis will be carried on on a forthcoming paper [53].

Figure 20 shows the average KL divergence between the true model and the inferred one for several color compression frequencies at low regularization ($\gamma_J = 1/B$). The dashed line is the KL divergence obtained with large regularization and without color compression for the same sample size. It is evident that in all cases the large regularization inference gives the best model. However, for all the sample size there is an optimal value of f_0, and especially at small sampling depth $B \leq 1000$ a large color compression leads to a very significant decreases of the KL divergence. For such cases the best color compression is around $f_0 = 0.5$, which reduces the model to a two-state model: the most common state and the grouped state, which gathers all the others. In the physics language such color compression reduces the Potts model to an Ising model. For large sample sizes ($B \sim 10^4, 10^5$), the best f_0 are around $0.01, 0.03$, while for stronger compressions the model loses its predictive power. However, in spite of the improvement due to color compression, the KL divergences do not reach the minimal values obtained at strong regularization (dashed lines in Fig. 20), showing that a good choice of the regularization is always essential.

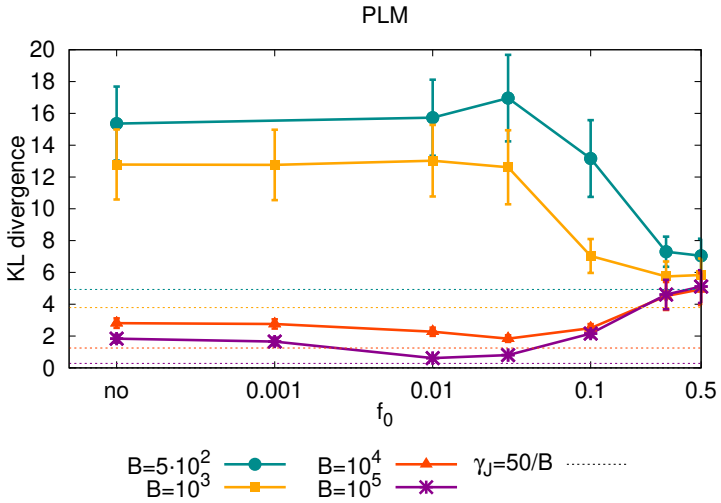


FIG. 20: KL divergence between the true model and the one inferred with $\gamma_J = 1/B$ averaged over 10 realizations for several sample sizes B at different color compression thresholds f_0 (full line). Error bars are standard deviations over the 10 realizations. Horizontal dashed lines are there for comparison and correspond to the KL values obtained with $\gamma_J = 50/B$ without color compression.

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

In section III A 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:

$$S_{ind} = \log \sum_{a=1}^{q} e^{h_i(a)} - \sum_{a=1}^{q} h_i(a)\, p_i(a) + \gamma_h \cdot \sum_{a=1}^{q} h_i(a)^2 \tag{29}$$

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

$$h_u^{\Gamma} = -Lw(\frac{1}{2\gamma_h}) \tag{30}$$

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,

$$h_u^{p} = \log(\frac{\alpha}{B}) \tag{31}$$

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 V 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 V 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 Figure 21 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

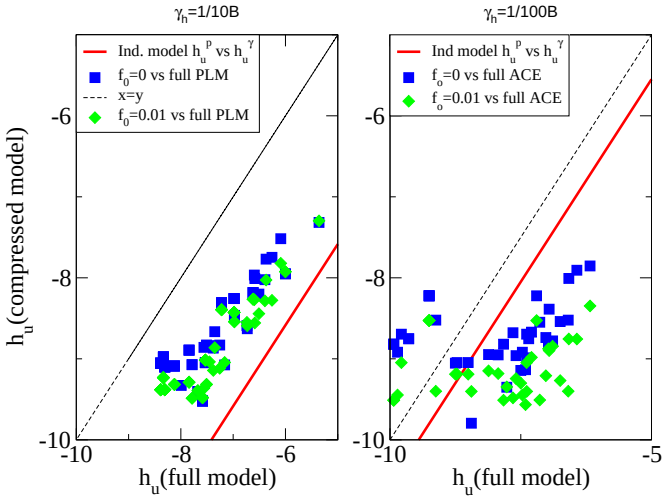


FIG. 21: Fields for the unseen Potts symbols for ER05 and $B = 1000$ in the non compressed model and the model with $f_0 = 0.01$ of Fig 12. Dotted line x=y Red Line: the shift given in the independent model for $\gamma_h = 1/10B$ as the one used with PLM.

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 V: 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

one for no color-compression for PLM and ACE (same fields of Fig. 12 and 13 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.

[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] H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. Bhat, H. Weissig, I. N. Shindyalov, and P. E. Bourne, "The protein data bank," *Nucleic Acids Research*, vol. 28, no. 1, pp. 235–242, 2000.

[11] A. Bateman, E. Birney, R. Durbin, S. R. Eddy, K. L. Howe, and E. L. Sonnhammer, "The pfam protein families database," *Nucleic acids research*, vol. 28, no. 1, pp. 263–266, 2000.

[12] 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.

[13] 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.

[14] 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.

[15] 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.

[16] 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.

[17] 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.

[18] 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.

[19] 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.

[20] 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.

[21] 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.

[22] 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.

[23] 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.

[24] 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.

[25] 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.

[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] 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.

[30] 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.

[31] 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.
[32] 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.
[33] 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.
[34] E. T. Jaynes, "On the rationale of maximum-entropy methods," *Proceedings of the IEEE*, vol. 70, no. 9, pp. 939–952, 1982.
[35] 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.
[36] 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.
[37] The first interval is $I/\tau < t < 1$, the second is for $I/\tau^2 < t < 1/\tau$ and so on. We have chosen here $\tau = 3.4$ (see command -trec on the GitHub site).
[38] R. Salakhutdinov, "Learning and evaluating boltzmann machines," tech. rep., UTML TR 2008?002, 2008.
[39] 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.
[40] S. Cocco, L. Posani, and R. Monasson, "Functional couplings from sequence and mutational data," *In Preparation*, 2019.
[41] 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.
[42] 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.
[43] 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.
[44] 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.
[45] 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.
[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] J. Cardy, *Scaling and renormalization in statistical physics*, vol. 5. Cambridge university press, 1996.
[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] J. Tubiana, S. Cocco, and R. Monasson, "Learning protein constitutive motifs from sequence data," *eLife*, vol. 8, p. e39397, 2019.
[52] C.-Y. Gao, H.-J. Zhou, and E. Aurell, "Correlation-compressed direct-coupling analysis," *Physical Review E*, vol. 98, no. 3, p. 032407, 2018.
[53] A. Fanthomme, S. Cocco, and R. Monasson, "Optimal regularizations for multivariate gaussian distributions," *In Preparation*, 2019.