Supplementary Material for: ACE: adaptive cluster expansion for maximum entropy graphical model inference

J. P. Barton,[1,2,*] E. De Leonardis,[3,4] A. Coucke,[3,5] and S. Cocco[4,†]

[1] *Departments of Chemical Engineering and Physics, Massachusetts Institute of Technology, Cambridge, MA 02139, USA*
[2] *Ragon Institute of Massachusetts General Hospital, Massachusetts Institute of Technology and Harvard, Cambridge, MA 02139, USA*
[3] *Computational and Quantitative Biology, UPMC, UMR 7238, Sorbonne Université, Paris, France*
[4] *Laboratoire de Physique Statistique de l'Ecole Normale Supérieure, CNRS, Ecole Normale Supérieure & Université P.&M. Curie, Paris, France*
[5] *Laboratoire de Physique Théorique de l'Ecole Normale Supérieure, CNRS, Ecole Normale Supérieure & Université P.&M. Curie, Paris, France*

I. METHODS

A. Cluster construction rules

In the original formulation of the cluster expansion algorithm, clusters of size $k+1$ were constructed when pairs of clusters of size k overlap by all but one site. Thus, in order to form the next set of clusters, $\mathcal{O}(N_c^2)$ operations are necessary, where N_c is the number of selected clusters of size k. For large systems and small values of the threshold, such pairwise comparisons can be computationally expensive.

To address this issue we developed a set of alternative cluster construction rules that can be executed quickly, even when the number of selected clusters is large. Rather than comparing pairs of clusters, we identify sets of sites that can be appended to existing clusters of size k to create larger clusters using a series of cutoffs, which we discuss below. We begin by counting the number of times n_i that each individual site i appears within the set of selected clusters. The number of times n_{ij} that each pair of sites i,j appear in one of the selected clusters is also recorded. The set of sites to be potentially added to the existing clusters is then chosen as $s_{\text{add}} = \{i | n_i \geq n_1^{\text{cut}}\}$. Once the set s_{add} is generated, we iterate through the list of selected clusters and create new candidate clusters of size $k+1$ by appending sites from s_{add}. We then consider pairs of sites i,j in the new cluster and check that $n_{ij} \geq n_2^{\text{cut}}$ for each pair. If this check passes, then we add the new cluster Γ to the list of potential new clusters, counting the number of times n_Γ that this cluster is formed as we iterate through the list of selected clusters. Finally, all candidate clusters with $n_\Gamma \geq n_c^{\text{cut}}$ are accepted. The computational complexity of this algorithm is $\mathcal{O}(N_c \times k^2)$, substantially faster than $\mathcal{O}(N_c^2)$ when the number of clusters is larger than the typical maximum cluster size (practically, $N_c \gtrsim 10$).

The rules described above depend on three cutoff values: n_1^{cut}, n_2^{cut}, and n_c^{cut}. We choose these cutoff values according to one of two conventions:

1. **Lax**: $n_1^{\text{cut}} = 1,\ n_2^{\text{cut}} = 0,\ n_c^{\text{cut}} = 2$. This is consistent with the rule that a new cluster is constructed if *any* pairs of selected clusters overlap by all but one site. Each site in the new cluster must thus appear in the list of selected clusters at least once, and the new candidate cluster must be formed at least two times (once from each one of the overlapping selected clusters), but each pair of sites need not be observed in the list of selected clusters.

2. **Strict**: $n_1^{\text{cut}} = k,\ n_2^{\text{cut}} = k-1,\ n_c^{\text{cut}} = k$. This rule is consistent with the requirement that *all* possible size k subclusters of a new size $k+1$ cluster to be added must be in the list of selected clusters. Here, each site in the new cluster must appear at least k times, and each pair at least $k-1$ times. The new candidate cluster must then be formed by each possible subcluster while iterating through the list.

We use the strict cluster construction rule by default because this focuses the cluster expansion toward sets of sites where significant interactions are clearly observed. Although the strict cluster construction rule can cause some significant clusters to be missed (at least until the value of the threshold is lowered), it results in the generation of

*jpbarton@mit.edu
†cocco@lps.ens.fr

much fewer clusters overall. The fraction of generated clusters that are selected is also much higher for the strict construction rule, leading to a more efficient expansion.

Gauge choice and evaluation of coupling values

As described in the main text, the Potts model is invariant under so-called gauge transformations, where

$$\begin{aligned} J_{ij}(a,b) &\to J_{ij}(a,b) + K_{ij}(a)\,, \\ h_i(a) &\to h_i(a) - \sum_{j\neq i} K_{ij}(a)\,, \end{aligned} \tag{1}$$

for any K. In addition, all the fields $h_i(a)$ at a site i can be uniformly shifted by a constant with no overall effect on the probability. Thus, the number of independent fields at each site i is $(q_i - 1)$ instead of q_i, and the number of independent couplings for each pair of sites is $(q_i - 1)(q_j - 1)$.

The gauge transformations in Eq. (1) can be exploited to choose a particular state of each variable (e.g. a particular amino acid at each site in a protein sequence) and set the field and all couplings corresponding to this state to zero. We refer to this as the gauge state for that variable. For protein sequences, common choices for the gauge state are the most common (consensus) amino acid at each site, gaps, or (in this work) the grouped state.

Note that, as has been noted previously [1], inferred couplings and fields are not invariant under the choice of the gauge due to the L_2-norm regularization, so it is important to fix a particular gauge to compare them. As far as the artificial models are concerned, we typically extract the parameters for a q-state Potts model, then to compare the inferred parameters to the true ones we need to have them in the same gauge. We refer to the latter as the *comparison gauge*. It can be different both from the original one and from the one chosen for the inference (in the following called the *inference gauge*). To do so we perform the following transformations, where the gauge states chosen for the comparison at site i are denoted by c_i:

$$\begin{aligned} J'_{ij}(a,b) &= J_{ij}(a,b) - J_{ij}(c_i,b) - J_{ij}(a,c_j) + J_{ij}(c_i,c_j)\,, \\ h'_i(a) &= h_i(a) - h_i(c_i) + \sum_{j\neq i} \left[J_{ij}(a,c_j) - J_{ij}(c_i,c_j)\right]\,. \end{aligned} \tag{2}$$

For contact map predictions, given a certain matrix J_{ij} we compress the extra information about interactions between different states using the Frobenius norm of the matrix. However before the Frobenius norm is computed, we have to put couplings in a different gauge in which the sum of the elements in each column and row is zero. It can be easily proved that this choice is the one minimizing the resulting Frobenius norm.

B. Entropy, couplings, fields, and gauge invariances for 2-variable clusters

Let's consider a system of 2 variables ($l = i, j$) with q_l states for each of them. The $q_i + q_j$ frequencies $p_l(a)$ and the $q_i \times q_j$ correlations $p_{ij}(a,b)$ have been sampled. The following conservations of the probabilities hold: $\sum_{a=1}^{q_l} p_l(a) = 1$, $\sum_{a=1}^{q_i} p_{ij}(a,b) = p_j(b)$, and $\sum_{a=1}^{q_i}\sum_{b=1}^{q_j} p_{ij}(a,b) = 1$. The probability of a configuration (a,b) for the two variables is expressed in the Potts model as:

$$p_{ij}(a,b) = e^{h_i(a)+h_j(b)+J_{ij}(a,b)} \tag{3}$$

with the partition function

$$Z = \sum_{a,b} e^{h_i(a)+h_j(b)+J_{ij}(a,b)} \equiv \sum_{a,b} p_{ij}(a,b) = 1 \tag{4}$$

The conditional probability of having b in position j given a in position i is in the Potts model $p(j,b|i,a) = e^{h_j(b)+J_{ij}(a,b)}$; by rewriting $p_{ij}(a,b) = p_i(a)p(j,b|i,a) = p_i(a)e^{h_j(b)+J_{ij}(a,b)}$ and comparing with Eq. (3) we obtain $p_i(a) = e^{h_i(a)}$ or:

$$h_i(a) = \log p_i(a)\,. \tag{5}$$

An analogous expression is obtained for $h_j(b)$. Substituting Eq. (5) for $h_i(a)$ and $h_j(b)$ in Eq. (3) we obtain

$$J_{ij}(a,b) = \log\left(\frac{p_{ij}(a,b)}{p_i(a)\,p_j(b)}\right) \tag{6}$$

It is easy to verify that the conservation equations for probabilities are satisfied by the aboves choice for the parameters $h_i(a), h_j(b)$ and $J_{ij}(a,b)$.

Note that the above equations for the couplings and fields are also obtained by deriving the minimal cross-entropy (see Eq. (4) in the main text) for a system of 2 variables with respect to the fields and couplings, which can be rewritten as

$$S(i,j) = \sum_{a,b} p_{ij}(a,b) \log\left(\frac{p_{ij}(a,b)}{p_i(a)p_j(b)}\right) + \sum_a p_i(a)\log p_i(a) + \sum_b p_j(b)\log p_j(b)\,. \tag{7}$$

Following Eq. (6) in the main text the 2-variable cluster contributions to the entropy are

$$\Delta S(i,j) = \sum_{a,b} p_{ij}(a,b) \log\left(\frac{p_{ij}(a,b)}{p_i(a)p_j(b)}\right). \tag{8}$$

The 2-variable cluster contributions to the couplings are $\Delta J_{ij}(a,b) = J_{ij}(a,b)$, defined in Eq. (6), and to the fields $\Delta h_i(a) = -\sum_b J_{ij}(a,b)p_j(b)$.

Due to the conservation laws we can fix a gauge for the fields and couplings (see Eq. (2)), *e.g.* by imposing that for each site i the field and the couplings for a chosen state c_i are equal to zero. In this way we have less parameters to infer in the model. The frequencies for this "gauge" state do not have to be recorded but can be derived from the other frequencies: $p_i(c) = 1 - \sum_{a\neq c} p_i(a)$, and $p_{ij}(a,c) = p_i(a) - \sum_{a'\neq c} p_{ij}(a,a')$. The gauge transformations such that $h'_i(c) = 0$, $J'_{ij}(c,b) = 0$ for the couplings and fields for 2-variable clusters are:

$$\begin{aligned} h'_i(a) &= \log p_i(a) - \log p_i(c_i) + \sum_{j=1}^{N} \left(J_{ij}(a,c_j) - J_{ij}(c_i,c_j)\right)\,, \\ J'_{ij}(a,b) &= \log p_{ij}(a,b) - \log p_{ij}(c_i,b) - \log p_{ij}(a,c_j) + \log p_{ij}(c_i,c_j)\,. \end{aligned} \tag{9}$$

C. Additional regularization schemes

1. *Sparse coupling constraint*

While the cluster expansion described above alleviates computational problems related to large system sizes, Potts models with large numbers of states (e.g. models of protein sequences, where 20 amino acids + 1 gap state may be allowed at each site) remain computationally difficult. The number of terms in the partition function is $\prod_i q_i$, thus even for a small cluster of sites a large number of terms may need to be summed if $q_i \gg 1$. To analyze such models, we employed L_0-regularization to enforce sparsity on the couplings combined with an efficient expansion of the partition function to exploit sparse coupling structure.

To understand how a sparse coupling structure can decrease computational costs, we first observe that, given the form of the energy (Eq. (1) of the main text), the partition function can be written in a particularly simple form in the case that all couplings are equal to zero:

$$Z = \prod_{i=1}^{N}\left(\sum_{a=1}^{q_i} e^{h_i(a)}\right). \tag{10}$$

Because the fields $h_i(a)$ at each site make independent contributions to the energy, the sum over all configurations can be expressed as a product of terms from each site. Eq. (10) thus requires only $\sum_{i=1}^{N} q_i$ operations to compute, rather than $\prod_{i=1}^{N} q_i$. In the general case, where all couplings are not equal to zero, we can partially factorize the partition function into a sum of independent and interacting factors. Let ν_i represent the set of states a at site i which have $J_{ij}(a,b) = 0$ for all j,b, and let μ_i denote the set of states that do not belong to ν_i (i.e. $J_{ij}(a,b) \neq 0$ for some j,b). We can then write a tree-like expansion the partition function as

$$Z = \left(\prod_{i=1}^{N} Z_i\right)\left[1 + \sum_{i=1}^{N}\sum_{a\in\mu_i} \frac{e^{h_i(a)}}{Z_i}\left[1 + \sum_{j=i+1}^{N}\sum_{b\in\mu_j}\frac{e^{h_j(b)+J_{ij}(a,b)}}{Z_j} \times \left[1 + \sum_{k=j+1}^{N}\sum_{c\in\mu_k}\frac{e^{h_k(c)+J_{ik}(a,c)+J_{jk}(b,c)}}{Z_k}\left[1+\ldots\right]\right]\right]\right]. \tag{11}$$

where $Z_i = \sum_{a\in\nu_i} \exp\left(h_i(a)\right)$. Eq. (11) can be efficiently computed when interactions are sparse, and in the simple case that $\mu_i = \emptyset$ for all i, it reduces simply to Eq. (10).

To enforce sparsity, we applied L_0-regularization for the couplings,

$$\Delta\ell = -\gamma_0 \sum_{i=1}^{N-1}\sum_{j=i+1}^{N}\sum_{a=1}^{q_i}\sum_{b=1}^{q_j} \|J_{ij}(a,b)\|_0 \,. \tag{12}$$

With this choice of regularization, couplings that do not increase the likelihood of the model (or equivalently, decrease the cross-entropy between the model and the data) by at least γ_0 are set to zero. This form of the regularization was implemented following the adaptive forward-backward algorithm presented in [2].

2. *Gauge invariant regularization of the couplings*

One can clearly show that the magnitude of the standard L_2-norm regularization term (Eq. (5) in the main text) is not gauge-invariant. That is, transformations of the form of Eq. (1) change the value of the regularization penalty. Thus, slightly different models can be inferred from the same data, and using the same regularization strength, if the gauge is fixed in different ways.

This dependence on the gauge choice can be avoided through the use of a gauge invariant regularization for couplings. Instead of an L_2-norm penalty on $J_{ij}(a,b)$, we instead introduce the penalty on a transformed coupling value

$$K_{ij}(a,b) = J_{ij}(a,b) - \frac{1}{q_j}\sum_{c=1}^{q_j} J_{ij}(a,c) - \frac{1}{q_i}\sum_{c=1}^{q_i} J_{ij}(c,b) + \frac{1}{q_i q_j}\sum_{c=1}^{q_i}\sum_{d=1}^{q_j} J_{ij}(c,d)\,. \tag{13}$$

One can then verify that $K_{ij}(a,b)$ is invariant under gauge transformations, and thus an L_2-norm regularization of the form

$$\gamma \sum_{i=1}^{N-1}\sum_{j=i+1}^{N}\sum_{a=1}^{q_i}\sum_{b=1}^{q_j} K_{ij}(a,b)^2 \tag{14}$$

does not depend on the choice of gauge.

D. Finite sampling error estimation

1. *Error on the frequencies and correlations*

The typical uncertainties of the 1- and 2-point frequencies can be determined simply from the covariance matrix,

$$\begin{aligned} \delta p_i &= \sqrt{\frac{1}{B}\,\chi_{i,i}} = \sqrt{\frac{\langle x_i\rangle_{\mathbf{J}}(1-\langle x_i\rangle_{\mathbf{J}})}{B}} \\ \delta p_{ij} &= \sqrt{\frac{1}{B}\,\chi_{ij,ij}} = \sqrt{\frac{\langle x_i x_j\rangle_{\mathbf{J}}(1-\langle x_i x_j\rangle_{\mathbf{J}})}{B}}\,. \end{aligned} \tag{15}$$

2. *Error on the inferred parameters*

The Hessian of the cross-entropy, χ, is the Fisher information matrix and it can be used to estimate the statistical fluctuations of the inferred parameters due to finite sampling. In the limit of large B, by the central limit theorem we know that the log-likelihood obeys a normal law centred on the minimum of $S_{\text{Potts}}(\mathbf{J}|\mathbf{p})$. Thus, given its covariance matrix $\frac{1}{B}\chi^{-1}$, we can define the errors on couplings and fields as follows:

$$\begin{aligned} \delta h_i(a) &= \sqrt{\frac{1}{B}(\chi^{-1})_{ia,ia}}\,, \\ \delta J_{ij}(a,b) &= \sqrt{\frac{1}{B}(\chi^{-1})_{iajb,iajb}}\,. \end{aligned} \tag{16}$$

To ensure that χ is positive definite we need to include a regularization term, as in Eq. (3) of the main text, before inverting the Hessian. Moreover, the inversion of χ is computationally infeasible for long sequences and for large q given that it has size $\left(qN + q^2(\frac{N(N-1)}{2})\right) \times \left(qN + q^2(\frac{N(N-1)}{2})\right)$, thus some approximate value for errors is needed in most biologically interesting cases (protein sequences typically have $N \sim 100$ and $q \gtrsim 10$). Below we derive a simple approximation for the error from the couplings and fields obtained only from pairs of variables.

3. *Two-variable approximation of the error the inferred parameters*

In the case that we consider just two variables alone, we have a simple approximation for the couplings and fields. We first regularize

$$p_i(a) \to p_i(a) + \frac{1}{B}\,,$$
$$p_{ij}(a,b) \to p_{ij}(a,b) + \frac{1}{B}\,.$$

We then obtain an approximate formula for the fields and couplings:

$$\begin{aligned} h_i(a) &= \log p_i(a) - \log p_i(c_i) + \sum_{j=1}^{N} \left[\log\left(\frac{p_{ij}(a,c_j)}{p_i(a)p_j(c_j)}\right) - \log\left(\frac{p_{ij}(c_i,c_j)}{p_i(c_i)p_j(c_j)}\right)\right]\,, \\ J_{ij}(a,b) &= \log p_{ij}(a,b) - \log p_{ij}(c_i,b) - \log p_{ij}(a,c_j) + \log p_{ij}(c_i,c_j)\,. \end{aligned} \tag{17}$$

The corresponding variances for the fields and couplings due to finite sampling are given by

$$\sigma_{h_i(a)} = (N-2)\,\frac{1-p_i(a)}{B\,p_i(a)} + (N-2)\,\frac{1-p_i(c_i)}{B\,p_i(c_i)} \sum_{j\neq i}\left(\frac{1-p_{ij}(a,c_j)}{B\,p_{ij}(a,c_j)} + \frac{1-p_{ij}(c_i,c_j)}{B\,p_{ij}(c_i,c_j)}\right)\,, \tag{18}$$

and standard deviations $dh_i(a) = \sqrt{\sigma_{h_i(a)}}$; and

$$\sigma_{J_{ij}(a,b)} = \frac{1-p_{ij}(a,b)}{B\,p_{ij}(a,b)} + \frac{1-p_{ij}(c_i,b)}{B\,p_{ij}(c_i,b)} + \frac{1-p_{ij}(a,c_j)}{B\,p_{ij}(a,c_j)} + \frac{1-p_{ij}(c_i,c_j)}{B\,p_{ij}(c_i,c_j)}\,. \tag{19}$$

and standard deviations $dJ_{ij}(a,b) = \sqrt{\sigma_{J_{ij}(ab)}}$.

E. Gaussian Inference

A faster approximate method to infer couplings and fields is using the Gaussian model. In this case the couplings are obtained as

$$J_{ij}(a,b) = c_{ij}^{-1}(ab)\,. \tag{20}$$

The above expression for the couplings is the same as the one obtained in the mean-field approximation. The fields are obtained as

$$h_i(a) = -\sum_{i\neq j,b} J_{ij}(a,b)\,p_j(b) - \sum_{b\neq a} J_{i,i}(a,b)\,p_i(b) + \frac{1}{2}J_{i,i}(a,a)\,(1-2\,p_i(a))\,. \tag{21}$$

These are different and generally give a better generative model than the one obtained through the mean-field approximation. Typically the frequencies and pairwise frequencies are pretreated with a large pseudo-count $\alpha = 0.5$:

$$p_i(a) \;\to\; \frac{1-\alpha}{q}\,p_i(a) + \frac{\alpha}{q}\,, \tag{22}$$
$$p_{ij}(a,b) \;\to\; \frac{1-\alpha}{q^2}\,p_{ij}(a,b) + \frac{\alpha}{q^2}\,. \tag{23}$$

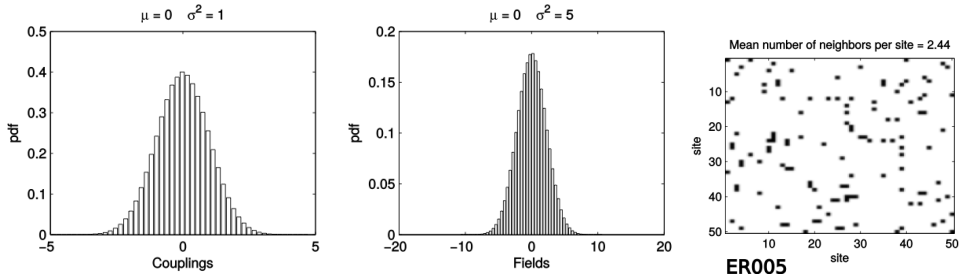

FIG. 1: **ER model parameters**. (Left, middle) Gaussian distributions from which parameters of the artificial model are chosen. For fields $\mu = 0$ and $\sigma^2 = 5$ while for couplings $\mu = 0$ and $\sigma^2 = 1$. (Right) "Contact map" for the ER05 model, where black squares represent interacting sites with non-zero couplings. If i and j interact then J_{ij} is a 21×21 matrix whose elements are chosen according to the above distributions. The maximum number of interacting sites for any site is 7. In total there are 61 interacting pairs of sites.

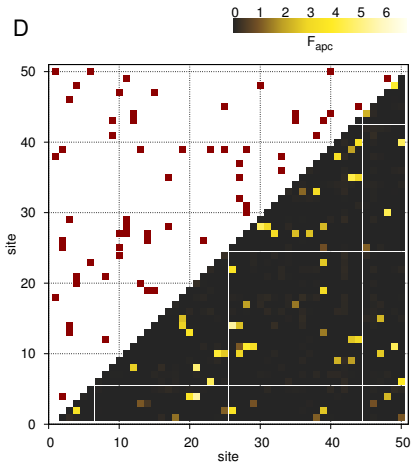

FIG. 2: **ER05 model inferred by ACE with $\mathbf{p_0 = 0.05}$**. Here we show the "contact map" obtained with the couplings inferred by ACE. The top-left triangle shows the top 61 predicted contact (i.e. interacting) pairs, all of which are correct. In the bottom-right triangle the full Frobenius norm with APC is displayed.

II. APPLICATIONS

A. Artificial data on Erdos-Renyi models

The network of interactions for the ER05 model studied here is shown in the contact map of Fig. 1. Coupling values for pairs of interacting sites are drawn according to a Gaussian distribution, shown in Fig. 1. Figure 2 shows that all contacts are correctly predicted from the 61 pairs of sites with the largest Frobenius norm of the inferred couplings, after the average product correction (APC).

Lattice-protein structure

In Fig. 3 we show the lattice protein fold of S_B to which the multi-sequence alignment is associated.

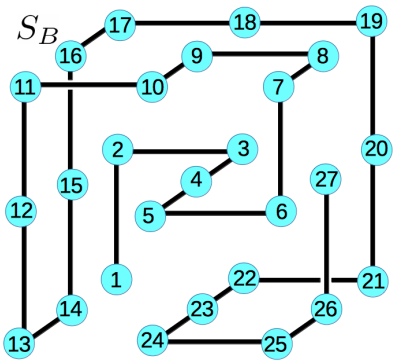

FIG. 3: **Structure of the lattice protein** S_B. The MSA analyzed is composed of sequences with a large probability to fold on the above lattice protein structure.

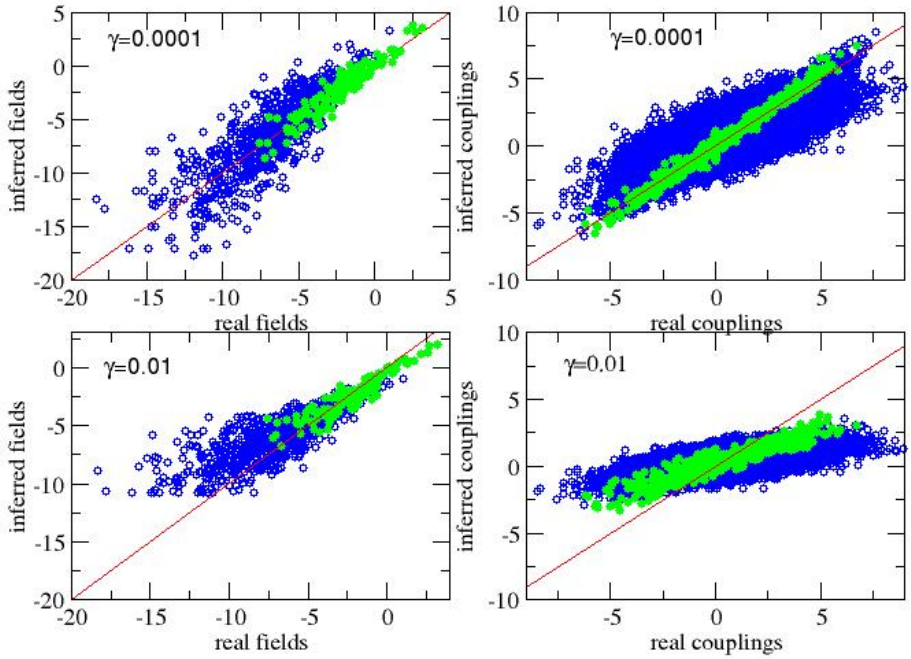

FIG. 4: **Pseudo-likelihood inference for the ER05 model**. Here we show scatter plots of the field and coupling parameters inferred by plmDCA compared against the true ones. The two values of the regularization we have used are $\gamma = 0.0001$ (top) and $\gamma = 0.01$ (bottom). Green points show the subset of parameters that correspond to explicitly modeled Potts states when using the compression $p_o = 0.05$.

Comparison with plmDCA for all data sets

The fields and couplings reconstructed for the ER005 data set with plmDCA using two values of the regularization $\gamma = 0.01$ and $\gamma = 0.0001$ are shown in Fig. 4. Note that for DCA and plmDCA we do not compress the number of states, thus more points are shown. The reconstruction of coupling parameters restricted to the ones selected in the ACE inference, after the Potts state compression with $p_0 = 0.05$, is quite accurate at the smallest regularization strength, but fields are less precisely inferred compared to ACE for both regularization strengths. Moreover, Fig. 4 and 5 show that Potts states that are not well sampled give less well inferred parameters. The presence of these poorly

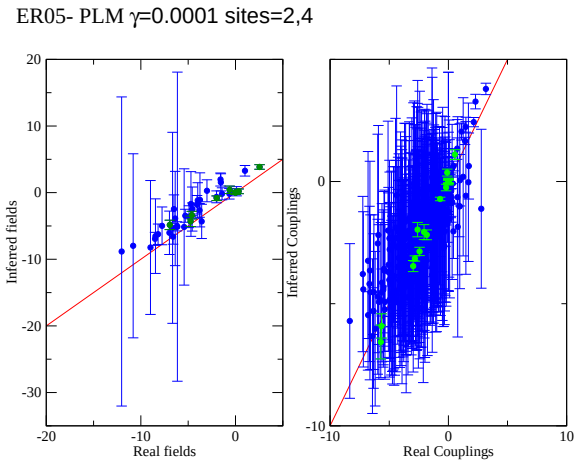

FIG. 5: **Pseudo-likelihood inference for the ER05 model**. Scatter plots of field and coupling parameters inferred by the plmDCA algorithm (with $\gamma = 0.0001$) shown against the true ones, restricted to the pair of interacting sites $(2, 4)$, and with statistical error bars for the inferred couplings. Green points show the subset of parameters that correspond to explicitly modeled Potts states when using the compression $p_o = 0.05$, which have smaller error bars.

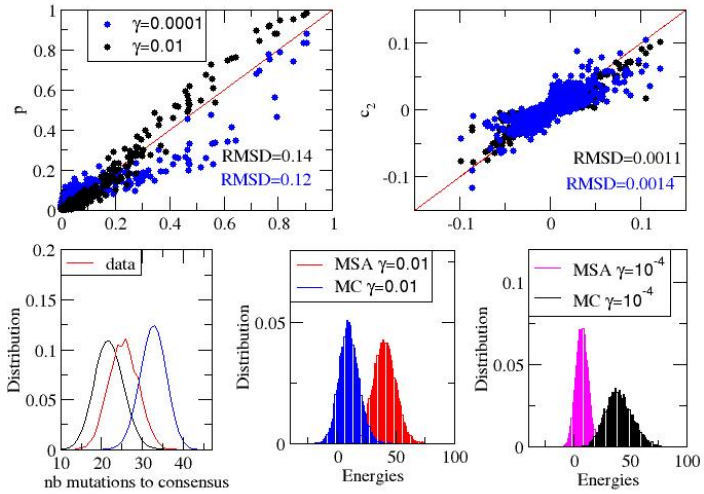

FIG. 6: **Pseudo-likelihood inference for the ER05 model**. Reconstruction of the statistics of the configurations by Monte Carlo sampling from the model inferred by plmDCA. Two values of the regularization are compared: $\gamma = 0.01$ and $\gamma = 10^{-4}$. The relative errors are $\epsilon_p = 9$ $(\gamma = 0.01)$, 18 $(\gamma = 10^{-4})$ and $\epsilon_{\max} = 40$ $(\gamma = 0.01)$, 233 $(\gamma = 10^{-4})$.

inferred parameters can also have an impact on the less accurate inference of fields and generative properties of the inferred model shown in Fig. 6. Reconstructions of statistics from the parameters inferred with plmDCA is shown for LP S_B (Fig. 7), the trypsin inhibitor PF00014 (Fig. 8), HIV p7 (Fig. 9), and the neural data (Fig. 10).

B. Gaussian (DCA) inference and generative tests for the ER05 model

The reconstruction of the frequencies, pairwise correlations, and distribution of the number of mutations with respect to the consensus sequence with the Gaussian model (DCA) for the ER05 data set is shown in Fig. 11. The

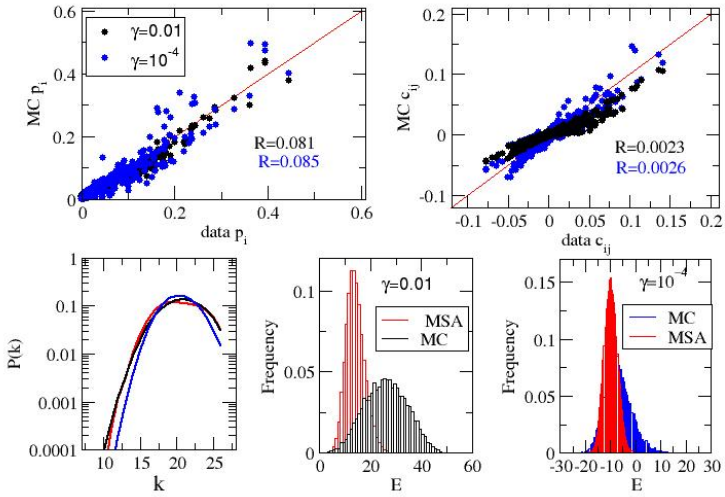

FIG. 7: **Pseudo-likelihood inference for the lattice protein** S_B. Reconstruction of the statistics of the configurations by Monte Carlo sampling from the model inferred by plmDCA. Two values of the regularization are compared: $\gamma = 0.01$ and $\gamma = 10^{-4}$. The relative errors are $\epsilon_p = 5$ ($\gamma = 0.01$), 2.4 ($\gamma = 10^{-4}$) and $\epsilon_{\max} = 5$ ($\gamma = 0.01$), 6 ($\gamma = 10^{-4}$).

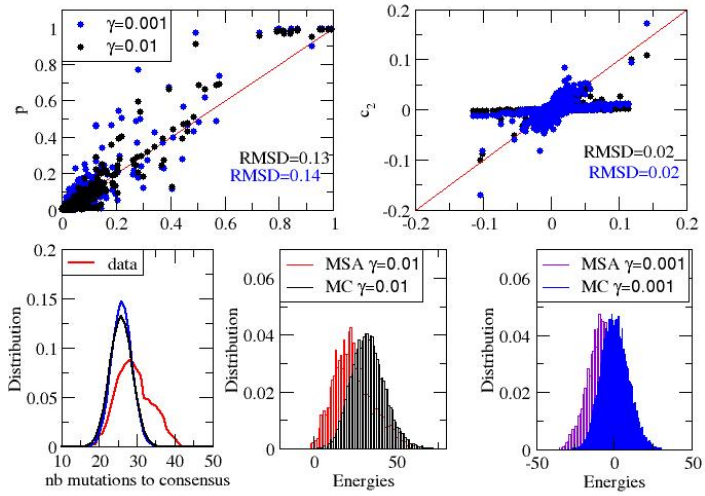

FIG. 8: **Pseudo-likelihood inference for the trypsin inhibitor PF00014**. Reconstruction of the statistics of the configurations by Monte Carlo sampling from the model inferred by plmDCA. Two values of the regularization are compared: $\gamma = 0.01$ and $\gamma = 10^{-3}$. The relative errors are $\epsilon_p = 4$ ($\gamma = 0.01$), 5 ($\gamma = 10^{-3}$) and $\epsilon_{\max} = 10$ ($\gamma = 0.01$), 21 ($\gamma = 10^{-3}$).

frequencies and pairwise correlation are not well reproduced and the distribution of configurations is strongly peaked around the consensus configuration.

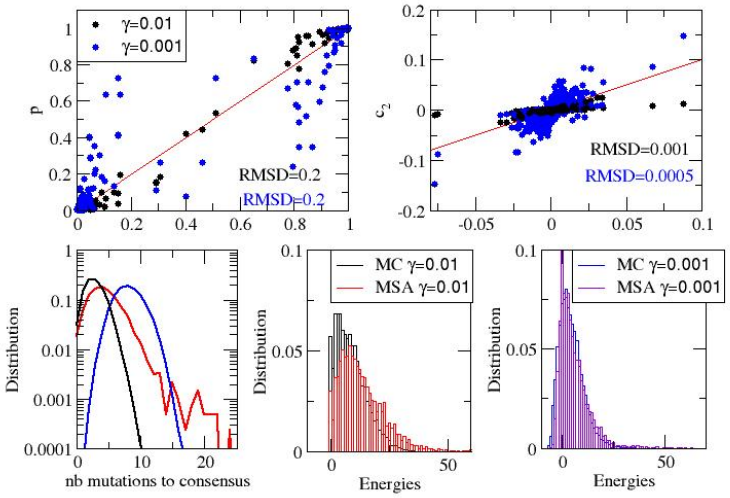

FIG. 9: **Pseudo-likelihood inference for the HIV protein p7**. Reconstruction of the statistics of the configurations by Monte Carlo sampling from the model inferred by plmDCA. Two values of the regularization are compared: $\gamma = 0.01$ and $\gamma = 10^{-3}$. The relative errors are $\epsilon_p = 2$ ($\gamma = 0.01$), 7 ($\gamma = 10^{-3}$) and $\epsilon_{\max} = 8$ ($\gamma = 0.01$), 90 ($\gamma = 10^{-3}$).

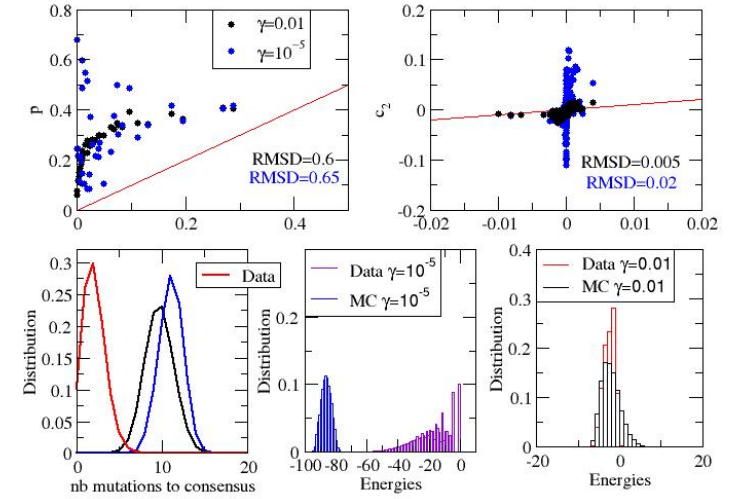

FIG. 10: **Pseudo-likelihood inference for the cortical data**. Reconstruction of the statistics of the configurations by Monte Carlo sampling from the model inferred by plmDCA. Two values of the regularization are compared: $\gamma = 0.01$ and $\gamma = 10^{-5}$. The relative errors are $\epsilon_p = 588 (\gamma = 0.01)$, 1700 ($\gamma = 10^{-5}$) and $\epsilon_{\max} = 2300$ ($\gamma = 0.01$), 18000 ($\gamma = 10^{-5}$).

C. Generative tests for ACE at very large regularization strength: application to PF00014

As underlined in the main text, the model inferred with ACE at very large regularization strength, $\gamma = 1$, gives very good contact predictions but it is not generative. The frequencies (Fig. 12) and high order non-connected correlations are quite well reproduced, but the connected correlations (Fig. 12) are very weak because the coupling parameters are largely overdamped.

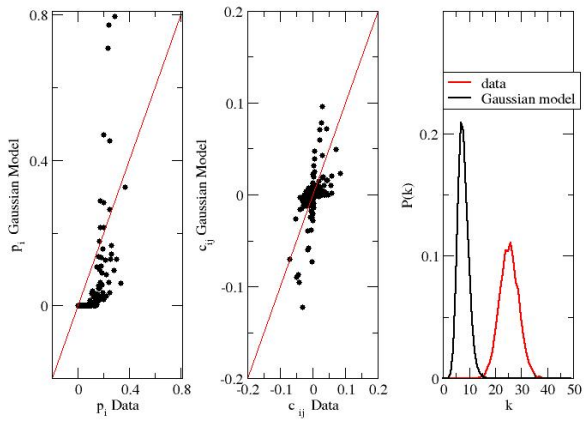

FIG. 11: **Gaussian (DCA) inference for the ER05 model**. Reconstruction of the statistics by Monte Carlo sampling from the model inferred using the Gaussian approximation.

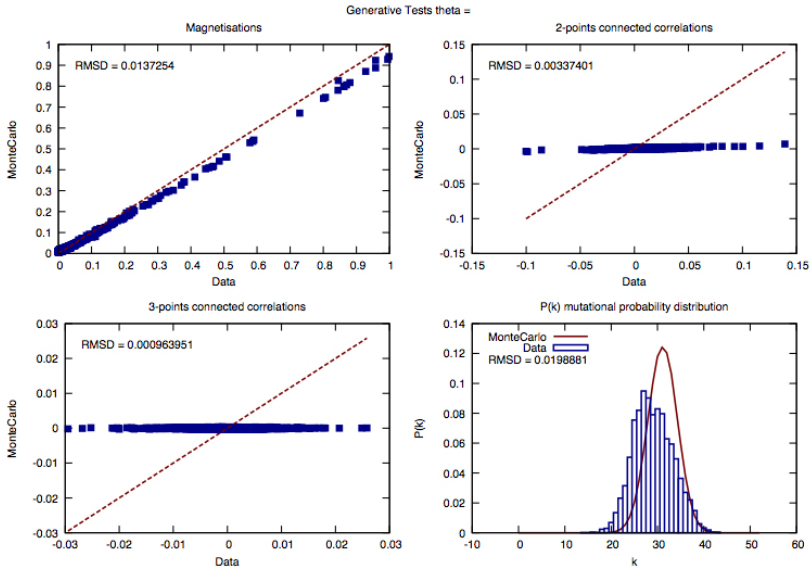

FIG. 12: **ACE inference with strong regularization ($\gamma = 1$) for PF00014**. Reconstruction of the statistics by Monte Carlo sampling from the model inferred by ACE with $\gamma = 1$.

D. Using ACE with knowledge of the contact map: application to PF00014

Here we start the inference procedure in ACE using a list of the set of two-site clusters corresponding to real contact pairs for PF00014. As shown in Fig. 13, clusters corresponding to the sites in contact are summed in the cluster entropy, even at threshold $t = 1$. When lowering the threshold the entropy reaches the curve obtained with the normal procedure as soon as the same clusters are summed. Fig. 14 shows the reconstruction of the statistics of the data from the parameters inferred at the small reduction $p_o = 0.005$ and for a large threshold $t = 0.04$, at which $\epsilon_{\max} = 3$.

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

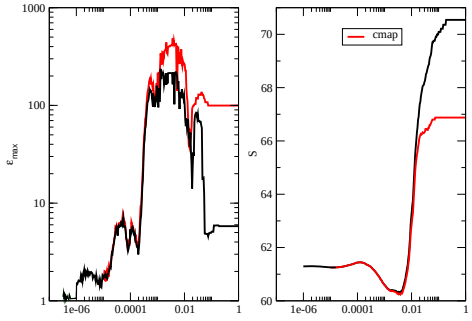

FIG. 13: **ACE inference for PF00014 starting from the list of contact pairs**. Maximal error and entropy S in the standard cluster expansion (black) are compared to those obtained when starting with knowledge of the contact map (red).

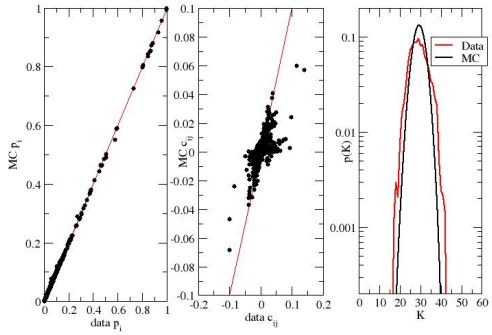

FIG. 14: **ACE inference with for PF00014 starting from the list of contact pairs with** $p_o = 0.005$. Reconstruction of statistics of the configurations from parameters inferred by ACE starting with knowledge of the contact map at a very large threshold $t = 0.04$.

[2] Zhang, T. (2009). Adaptive forward-backward greedy algorithm for sparse learning with linear models. In D. Koller, D. Schuurmans, Y. Bengio, and L. Bottou, editors, *Advances in Neural Information Processing Systems 21*, pages 1921–1928. Curran Associates, Inc.