Collective-variable selection and generative Hopfield-Potts models for protein-sequence families

Kai Shimagaki[1] and Martin Weigt[1]

[1]*Sorbonne Université, CNRS, Institut de Biologie Paris-Seine, Laboratoire de Biologie Computationnelle et Quantitative – LCQB Paris, France*

(Dated: May 27, 2019)

Statistical models for families of evolutionary related proteins have recently gained interest: in particular pairwise Potts models, as those inferred by the Direct-Coupling Analysis, have been able to extract information about the three-dimensional structure of folded proteins, and about the effect of amino-acid substitutions in proteins. These models are typically requested to reproduce the one- and two-point statistics of the amino-acid usage in these protein families, *i.e.* the so-called residue conservation and covariation statistics. Pairwise Potts models are the maximum-entropy models achieving this. While being successful, these models depend on huge numbers of ad hoc introduced parameters, which have to be estimated from finite amount of data and whose biophysical interpretation remains unclear. Here we propose an approach to parameter reduction, which is based on collective-variable selection. It naturally leads to the formulation of statistical sequence models in terms of Hopfield-Potts models, with the Hopfield patterns being the collective variables. These models can be accurately inferred using a mapping to restricted Boltzmann machines and persistent contrastive divergence. We show that, when applied to protein data, even 20-40 patterns are sufficient to obtain statistically generative models. The Hopfield patterns are interpretable in terms of sequence motifs and may be used to clusterize amino-acid sequences into functional subfamilies. However, the distributed collective nature of these motifs intrinsically limits the ability of Hopfield-Potts models in predicting contact maps, showing the necessity of developing models going beyond the Hopfield-Potts models discussed here.

PACS numbers: 87.14.E- Proteins, 87.15.Qt Sequence analysis, 02.50.-r Probability theory, stochastic processes, and statistics

I. INTRODUCTION

Thanks to important technological advances, including in particular next-generation sequencing, biology is currently undergoing a deep transformation towards a data-rich science. As an example, the number of available protein sequences deposited in the Uniprot database was about 1 million in 2004, crossed 10 millions in 2010, and 100 millions in 2018, despite an important reorganization of the database in 2015 to reduce redundancies and thus limit database size [1]. On the contrary, proteins with detailled experimental knowledge are contained in the SwissProt sub-database of Uniprot. While their number, currently close to 560 000, remained almost constant over the last decade, the knowledge about these selected proteins has been continuously extended and updated.

This fastly growing wealth of data is presenting both a challenge and an opportunity for data-driven modeling approaches. It is a challenge, because there are less than 0.5% of all known protein sequences, for which at least some knowledge going beyond sequence is available. Applicability of standard supervised machine-learning approaches is thus frequently limited. However, more importantly, it is an opportunity since collections of protein sequences are not large sets of unrelated random sequences, but contain structured, functional proteins resulting from natural evolution.

In particular, protein sequences can be classified into so-called *homologous protein families* [2]. Each family contains protein sequences, which are believed to share common ancestry in evolution. Such homologous sequences typically show very similar three-dimensional folded structures, and closely related biological functions. Put simply, they can be seen as equivalent proteins in different species, or in different pathways of the same species. Despite this high level of structural and functional conservation, homologous proteins may differ in more than 70-80% of their amino-acids. Detecting homology between well-studied and experimentally uncharacterized proteins [3, 4] is therefore the currently most important means for computational sequence annotation, including protein-structure prediction by homology modeling [5, 6].

Going beyond knowledge transfer, the observable sequence variability between homologous proteins contains on its own important information about the evolutionary constraints acting on proteins to conserve their structure and function [7]. Typically very few random mutations do actually destabilize proteins or interrupt their function. Some positions need to be highly conserved, while others are permissive for multiple mutations. Observing sequence variability across entire homologous protein families, and relating them to protein structure, function, and evolution, is therefore an important task [8].

Over the last years, *inverse statistical physics* [9] has played an increasing role in solving this task. Methods like

Direct-Coupling Analysis (DCA) [10, 11] or related approaches [12, 13] allow for predicting protein structure [14, 15], mutational effects [16–18] and protein-protein interactions [19]. However, many of these methods depend on huge numbers of typically *ad hoc* introduced parameters, making these methods data-hungry and susceptible to overfitting effects.

In this paper, we describe an attempt to substantially reduce the amount of parameters, and to select them systematically using sequence data. Despite this parameter reduction, we aim at so-called *generative* statistical models: samples drawn from these models should be statistically similar to the real data, even if similarity is evaluated using statistical measures, which were not used to infer the model from data.

To this aim, we first review in Sec. II some important points about protein-sequence data, maximum-entropy models of these data in general, and profile and DCA models in particular. In Sec. III, we introduce a way for rational collective variable selection, which generalizes maximum-entropy modeling. The resulting Hopfield-Potts models are mapped to Restricted Boltzmann Machines (recently introduced independently for proteins in [20]) in Sec. IV to enable efficient model inference and interpretation of the model parameters. Sec. V is dedicated to the application of this scheme to some exemplary protein families. The Conclusion and Outlook in Sec. VI is followed by some technical appendices.

II. A SHORT SUMMARY: SEQUENCE FAMILIES, MAXENT MODELS AND DCA

To put our work into the right context, we need to review shortly some published results about the statistical modeling of protein families. After introducing the data format, we summarize the maximum-entropy approach typically used to justify the use of Boltzmann distributions for protein families, together with some important shortcomings of this approach. Next we give a concise overview over two different types of maximum-entropy models – profile models and direct-coupling analysis – which are currently used for protein sequences. In all cases we discuss the strengths and limitations, which have motivated our current work.

A. Sequence data

Before discussing modeling strategies, we need to properly define what type of data is used. Sequences of homologous proteins are used in the form of *multiple-sequence alignments* (MSA), *i.e.* rectangular matrices $(A_i^m)_{i=1,...,L}^{m=1,...,M}$. Each of the rows $m = 1, ..., M$ of this matrix contains one aligned protein sequence $\underline{A}^m = (A_1^m, ..., A_L^m)$ of length L. In the context of MSA, L is also called the alignment width, M its depth. Entries in the matrix come from the alphabet $\mathcal{A} = \{-, A, C, ..., Y\}$ containing the 20 natural amino acids and the alignment gap "–". Throughout this paper, the size of the alignment will be denoted by $q = 21$. Practically we will use a numerical version of the alphabet, denoted by $\{1, ..., q\}$, but we have to keep in mind that variables are categorical variables, *i.e.* there is no linear order associated to these numerical values.

The Pfam database [2] currently (release 32.0) lists almost 18,000 protein families. Statistical modeling is most successful for large families, which contain between 10^3 and 10^6 sequences. Typical lengths span the range $L = 30-500$.

B. Maximum-entropy modeling

The aim of statistical modeling is to represent each protein family by a function $P(\underline{A})$, which assigns a probability to each sequence $\underline{A} \in \mathcal{A}^L$. Obviously the number of sequences even in the largest MSA is much smaller than the number $q^L - 1$ of a priori independent parameters characterizing P. So we have to use clever parameterizations for these models.

A commonly used strategy is the maximum-entropy (MaxEnt) approach [21]. It start from any number p of observables,

$$\mathcal{O}^\mu : \mathcal{A}^L \to \mathbb{R} \ , \qquad \mu = 1, ..., p, \tag{1}$$

which assign real numbers to each sequence. Only the values of these observables for the sequences in the MSA $(\underline{A}^m)$ go into the MaxEnt models. More precisely, we require the model to reproduce the mean of each observable over the data:

$$\forall \mu = 1, ...p \ : \qquad \sum_{\underline{A} \in \mathcal{A}^L} P(\underline{A}) \mathcal{O}^\mu(\underline{A}) = \frac{1}{M} \sum_{m=1}^{M} \mathcal{O}^\mu(\underline{A}^m) \ . \tag{2}$$

In a more compact notation, we write $\langle \mathcal{O}^\mu \rangle_P = \langle \mathcal{O}^\mu \rangle_{\rm MSA}$. Besides this consistency with the data, the model should be as unconstrained as possible. Its entropy has therefore to be maximized,

$$-\sum_{\underline{A}\in\mathcal{A}^L} P(\underline{A}) \log P(\underline{A}) \longrightarrow \max \ . \tag{3}$$

Imposing the constraints Eq. (2) via Lagrange multipliers $\lambda_\mu, \mu = 1, ..., p$, we immediately find that $P(\underline{A})$ assumes a Boltzmann-like exponential form

$$P(\underline{A}) = \frac{1}{Z} \exp\left\{ \sum_{\mu=1}^{p} \lambda_\mu \mathcal{O}^\mu(\underline{A}) \right\} \ . \tag{4}$$

Model inference consists in fitting the Lagrange multipliers such that Eqs. (2) are satisfied. The partition function Z guarantees normalization of P.

MaxEnt relates observables and the analytical form of the probability distribution, but it does not provide any rule how to select observables. Frequently prior knowledge is used to decide, which observables are "important" and which not. More systematic approaches therefore have to address at least the following two questions:

- Are the selected observables *sufficient*? In the best case, model P becomes *generative*, *i.e.* sequences $\underline{A}$ sampled from P are statistically indistinguishable from the natural sequences in the MSA ($\underline{A}^m$) used for model learning. While this is hard to test in full generality, we can select observables *not* used in the construction of the model, and check if their statistics coincide in the model and the input data.

- Are the selected observables *necessary*? Would it be possible to construct a parameter-reduced, thus more parsimonious model of same quality? This question is very important due to at least two reasons: (a) the most parsimonious model would allow for identifying a minimal set of evolutionary constraints acting on proteins, and thus offer deep insight into protein evolution; and (b) a reduced number of parameters would allow to reduce overfitting effects, which result from the limited availability of data.

While there has been promising progress in the first question, cf. the next two subsections, our work attempts to approach both questions simultaneously, thereby going beyond standard MaxEnt modeling.

To facilitate the further discussion, two important technical points have to be mentioned here. First, MaxEnt leads to a family of so-called exponential models, where the exponent in Eq. (4) is *linear* in the Lagrange multipliers λ_μ, which parameterize the family. Second, MaxEnt is intimately related to maximum likelihood. When we postulate Eq. (4) for the mathematical form of model $P(\underline{A})$, and when we maximize the log-likelihood

$$\mathcal{L}(\{\lambda_\mu\} \,|\, (A_i^m)) = \sum_{m=1}^{M} \log P(\underline{A}^m) \tag{5}$$

with respect to the parameters $\lambda_\mu, \mu = 1, ..., P$, we rediscover Eqs. (2) as the stationarity condition. The particular form of $P(\underline{A})$ guarantees that the likelihood is convex, having only a unique maximum.

C. Profile models

The most successful approach in statistical modeling of biological sequences are probably *profile models* [22], which consider each MSA column (*i.e.* each position in the sequence) independently. The corresponding observables are simply $\mathcal{O}^{ia}(\underline{A}) = \delta_{A_i,a}$ for all positions $i = 1, ..., L$ and all amino-acid letters $a \in \mathcal{A}$, with δ being the standard Kronecker symbol. These observables thus just ask if in a sequence $\underline{A}$, amino acid a is present in position i. Their statistics in the MSA is thus characterized by the fraction

$$f_i(a) = \frac{1}{M} \sum_{m=1}^{M} \delta_{A_i^m,a} \tag{6}$$

of sequences having amino acid a in position i. Consistency of model and data requires marginal single-site distributions of P to coincide with the f_i,

$$\forall i = 1, ..., L, \forall A_i \in \mathcal{A} \ : \qquad \sum_{\{A_j | j \neq i\}} P(\underline{A}) = f_i(A_i) \ . \tag{7}$$

The MaxEnt model results as $P(\underline{A}) = \prod_i f_i(A_i)$, which can be written as a factorized Boltzmann distribution

$$P(\underline{A}) = \frac{1}{Z} \exp\left\{ \sum_i h_i(A_i) \right\} \ , \tag{8}$$

where the local fields equal $h_i(a) = \log f_i(a)$. Pseudocounts or regularization can be used to avoid infinite parameters for amino acids, which are not observed in some MSA column.

Profile models reproduce the so-called *conservation* statistics of an MSA, *i.e.* the heterogenous usage of amino acids in the different positions of the sequence. Conservation of a single or few amino-acids in a column of the MSA is typically an indication of an important functional or structural role of that position. Profile models, frequently in their generalization to profile Hidden Markov Models [3, 4, 23], are used for detecting homology of new sequences to protein families, for aligning multiple sequences, and - using the conserved structural and functional characteristics of protein families - indirectly for the computational annotation of experimentally uncharacterized amino-acid sequences. They are in fact at the methodological basis of the generation of the MSA used here.

Despite their importance in biological-sequence analysis, profile models are not generative. Biological sequences show significant correlations in the usage of amino acids in different positions, which are said to *coevolve* [7]. Due to their factorized nature, profile models are not able to reproduce these correlations, and larger sets of observables have to be used in potentially generative sequence models.

D. Direct-Coupling Analysis

The Direct-Coupling Analysis (DCA) [10, 11] therefore includes also pairwise correlations into the modeling. The statistical model $P(\underline{A})$ is not only required to reproduce the amino-acid usage of single MSA columns, but also the fraction $f_{ij}(a,b)$ of sequences having simultaneously amino acid a in position i, and amino acid b in position j, for all $a, b \in \mathcal{A}$ and all $1 \leq i < j \leq L$:

$$\begin{aligned} f_{ij}(a,b) &= \frac{1}{M} \sum_{m=1}^{M} \delta_{A_i^m,a}\, \delta_{A_j^m,b} \\ &= \sum_{\underline{A} \in \mathcal{A}^L} P(\underline{A})\, \delta_{A_i,a}\, \delta_{A_j,b} \ . \end{aligned} \tag{9}$$

The corresponding observables $\delta_{A_i,a}\, \delta_{A_j,b}$ are thus products of pairs of observables used in profile models.

According to the general MaxEnt scheme described before, DCA leads to a generalized q-states Potts model

$$P(\underline{A}) = \frac{1}{Z} \exp\left\{ \sum_{i<j} J_{ij}(A_i, A_j) + \sum_i h_i(A_i) \right\} \tag{10}$$

with heterogeneous pairwise couplings $J_{ij}(a,b)$ and local fields $h_i(a)$. The inference of parameters becomes computationally hard, since the computation of the marginal distributions in Eq. (9) requires to sum over $O(q^L)$ sequences. Many approximation schemes have been proposed, including message-passing [10], mean field [11], Gaussian [13, 24], pseudo-likelihood maximization [12, 25]. DCA and related global inference techniques have found widespread applications in the prediction of protein structures, of protein-protein interactions and of mutational effects, demonstrating that amino-acid covariation as captured by the f_{ij} contains biologically valuable information.

While these approximate inference schemes do not lead to generative models – not even the f_{ij} are perfectly reproduced due to their approximate character – recently very precise but time-extensive inference schemes based on Boltzmann-machine learning have been proposed [26–29]. Astonishingly, these models do not only reproduce the fitted one- and two-column statistics of the input MSA: also non-fitted characteristics like the three-point statistics $f_{ijk}(a,b,c)$ or the clustered organization of sequences in sequence space are reproduced. These observations strongly suggest that pairwise Potts models as infered via DCA are generative models, *i.e.* that the observables used in DCA – amino-acid occurrence in single positions and in position pairs – is actually defining a (close to) sufficient statistics. In a seminal experimental work [30], the importance of respecting pairwise correlations in amino-acid usage in generating small artificial but folding protein sequences was shown.

However, DCA uses an enormous amount of parameters. There are independent couplings for each pair of positions and amino acids. In case of a protein of limited length $L = 200$, the total number of parameters is close to 10^8. Very few of these parameters are interpretable in terms of, *e.g.*, contacts between positions in the three-dimensional protein

fold. We would therefore expect that not all of these observables are really important to model protein sequences statistically. On the contrary, given the limited size ($M = 10^3 - 10^4$) of most input MSA, the large number of parameters makes overfitting likely, and quite strong regularization necessary. It would therefore be important to devise parameter-reduced models [31].

III. COLLECTIVE VARIABLE SELECTION AND THE HOPFIELD-POTTS MODEL

Seen the importance of amino-acid conservation in proteins, and of profile models in computational sequence analysis, we keep Eqs. (7), which link the single-site marginals of $P(\underline{A})$ directly to the amino-acid frequencies $f_i(a)$ in single MSA columns. Further more, we assume that the important observables for our protein sequence ensemble can be represented as *collective variables*

$$\mathcal{O}^\mu(\underline{A}) = \sum_i \omega_i^\mu(A_i) \ , \quad \mu = 1, ..., p, \tag{11}$$

which are linear additive combinations of single-site terms. In sequence bioinformatics, these observables are also known as *sequence motifs*, or *position-specific scoring / weight matrices* (PSSM), cf. [32, 33]. Let us assume for a moment that these observables, or more specifically the corresponding ω-matrices, are known. We will address their selection later.

For any model P reproducing the sequence profile, *i.e.* for any model fulfilling Eqs. (7), also the ensemble average of the $\mathcal{O}^\mu$ is given,

$$\sum_{\underline{A}} P(\underline{A})\mathcal{O}^\mu(\underline{A}) = \sum_{i,a} \omega_i^\mu(a) f_i(a) \ . \tag{12}$$

The mean of these observables therefore does not contain further information about the MSA statistics than the profile itself. The key step is to consider also the variance, or the second moment,

$$\frac{1}{M}\sum_m \left[\mathcal{O}^\mu(\underline{A}^m)\right]^2 = \sum_{i,j,a,b} \omega_i^\mu(a)\omega_j^\mu(b) f_{ij}(a,b) \tag{13}$$

as a distinct feature characterizing the sequence variability in the MSA, which has to be reproduced by the statistical model $P(\underline{A})$. This second moment actually depends on the f_{ij}, which were introduced in DCA to account for the correlated amino-acid usage in pairs of positions.

The importance of fixing this second moment becomes clear in a very simple example: Consider only two positions $\{1, 2\}$ and two possible letters $\{A, B\}$, which are allowed in these two positions. Let us assume further that these two letters are equiprobable in these two positions, *i.e.* $f_1(A) = f_1(B) = f_2(A) = f_2(B) = 1/2$. Assume further the PSSM to be given as $\omega_1(A) = \omega_2(A) = 1/2$, $\omega_1(B) = \omega_2(B) = -1/2$. In this case, the mean of $\mathcal{O}$ equals zero. We further consider two cases:

- *Uncorrelated positions:* In this case, all words AA, AB, BA, BB are equiprobable. The second moment of $\mathcal{O}$ thus equals 1/2.
- *Correlated positions:* As a strongly correlated example, only the words AA, BB are assumed to be allowed. The second moment of $\mathcal{O}$ thus equals 1.

We conclude that an increased second moment (or variance) of these additive observables with respect to the uncorrelated case corresponds to the preference of combinations of letters or entire words, which are the before mentioned sequence motifs.

Including therefore these second moments as conditions into the MaxEnt modeling, our statistical model takes the shape

$$P\left(\underline{A} \,|\, \{\lambda_\mu, h_i(a), \omega_i^\mu(a)\}\right) = \frac{1}{Z}\exp\left\{\sum_{\mu=1}^{p}\lambda_\mu \sum_{i,j=1}^{L}\omega_i^\mu(A_i)\omega_j^\mu(A_j) + \sum_{i=1}^{L} h_i(A_i)\right\} \tag{14}$$

with Lagrange multipliers $\lambda_\mu, \mu = 1, ..., p$, imposing means (13) to be reproduced by the model, and $h_i(a), i = 1, ..., L, a \in \mathcal{A}$, to impose Eqs. (7).

A. The Hopfield-Potts model: from MaxEnt to collective-variable selection

As mentioned before, an important limitation of MaxEnt models is that they assume certain observables to be reproduced, but they do not offer any strategy, how these observables have to be selected. In the case of Eq. (14), this accounts in particular to optimizing the values of the Langrange parameters λ_μ to match the ensemble averages over $P(\underline{A})$ with the sample averages Eq. (13) over the input MSA. As mentioned before, this corresponds also to *maximizing the log-likelihood* of these parameters given MSA and observables,

$$\mathcal{L}(\{\lambda_\mu, h_i(a)\} \,|\, (A_i^m), \{\omega_i^\mu(a)\}) = \sum_{m=1}^{M} \log P(\underline{A}^m \,|\, \{\lambda_\mu, h_i(a), \omega_i^\mu(a)\}) \ . \tag{15}$$

The important, even if quite straight forward *step from MaxEnt modeling to Collective Variable Selection* (CVS) is to optimize the likelihood also over the choice of possible PSSM as parameters. To remove degeneracies, we absorbe the Lagrange multipliers λ_μ into the PSSM ω^μ, by introducing

$$\xi_i^\mu(a) = \sqrt{\lambda_\mu}\, \omega_i^\mu(a), \quad i = 1, ..., L;\ \mu = 1, ..., p;\ a \in \mathcal{A} \ . \tag{16}$$

The model in Eq. (14) thus slightly simplifies into

$$P(\underline{A} \,|\, \{h_i(a), \xi_i^\mu(a)\}) = \frac{1}{Z} \exp\left\{ \sum_{\mu=1}^{p} \sum_{i,j=1}^{L} \xi_i^\mu(A_i)\xi_j^\mu(A_j) + \sum_{i=1}^{L} h_i(A_i) \right\} \ , \tag{17}$$

with parameters, which have to be estimated by maximum likelihood:

$$\{\hat{h}_i(a), \hat{\xi}_i^\mu(a)\} = \mathrm{argmax}_{\{h_i(a), \xi_i^\mu(a)\}} \sum_{m=1}^{M} \log P(\underline{A}^m \,|\, \{h_i(a), \xi_i^\mu(a)\}) \ . \tag{18}$$

Our model is therefore the standard *Hopfield-Potts model*, which has been introduced in [31] in a mean-field treatment. The mean-field approximation has allowed to relate the Hopfield-Potts patterns ξ^μ to the eigenvectors of the Pearson-correlation matrix of the MSA, and their likelihood contributions to a function of the corresponding eigenvalues. However, the mean-field treatment leads to a non-generative model, which does not even reproduce precisely the single-position frequencies f_i.

The model contains now an exponent, which is non-linear in the parameters ξ^μ. As a consequence, the likelihood is not convex any more, and possibly many local likelihood maxima exist. This is also reflected by the fact that any p-dimensional orthogonal transformation of the ξ^μ leafs the probability distribution $P(\underline{A})$ invariant, thus leading to an equivalent model.

IV. INFERENCE AND INTERPRETATION OF HOPFIELD-POTTS MODELS

A. The Hopfield-Potts model as a Restricted Boltzmann Machine

The question how many and which patterns are needed for generative modeling therefore cannot be answered properly within the mean-field approach. We therefore propose a more accurate inference scheme based on *Restricted Boltzmann Machine* (RBM) learning [34, 35]. To this aim, we first perform p Hubbard-Stratonovich transformations to linearize the exponential in the ξ^μ,

$$P(\underline{A}) = \frac{1}{\tilde{Z}} \int_{\mathbb{R}^p} \prod_{\mu=1}^{p} dx^\mu \exp\left\{ \sum_{i,\mu} x^\mu \xi_i^\mu(A_i) + \sum_i h_i(A_i) - \frac{1}{2} \sum_\mu (x^\mu)^2 \right\} \tag{19}$$

with $\tilde{Z}$ containing the normalizations both of the Gaussian integrals over the new variables x^μ, and the partition function of Eq. (14). The distribution $P(\underline{A})$ can thus be understood as a marginal distribution of

$$P(\underline{A}, \underline{x}) = \frac{1}{\tilde{Z}} \exp\left\{ \sum_{i,\mu} x^\mu \xi_i^\mu(A_i) + \sum_i h_i(A_i) - \frac{1}{2} \sum_\mu (x^\mu)^2 \right\} \ , \tag{20}$$

which depends on the so-called *visible variables* $\underline{A} = (A_1, ..., A_L)$ and the *hidden (or latent) variables* $\underline{x} = (x^1, ..., x^p)$. It takes the form or a particular RBM, with a Gaussian potential for the x^μ: The important point is that couplings in the RBM form a bipartite graph between visible and hidden variables, cf. Fig. 1. RBM may have more general potentials $u_\mu(x^\mu)$ confining the values of the new random variables x^μ. In this work and in difference to [20], due to the relation with Hopfield-Potts models and DCA, we restrict all discussions to Gaussian potentials.

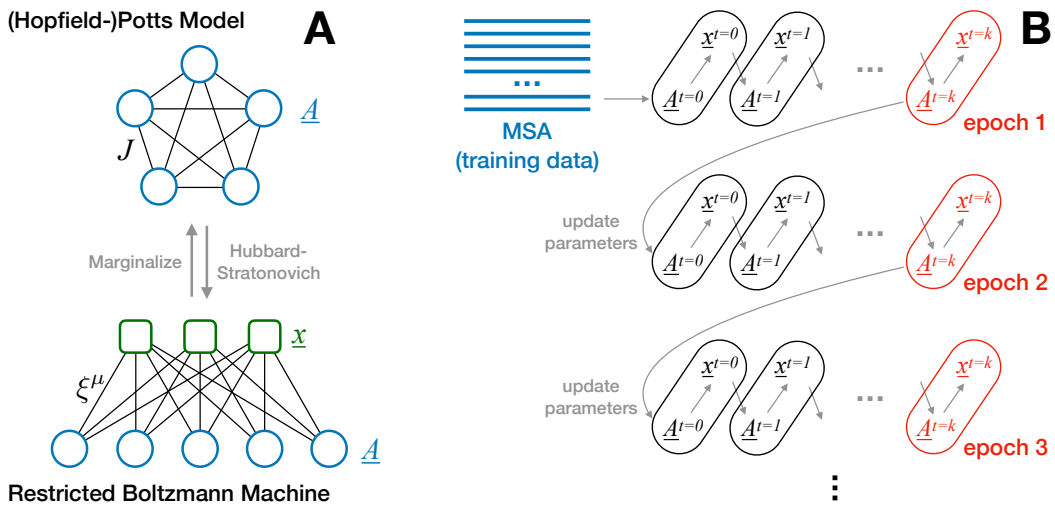

FIG. 1: **Panel A** represents the (Hopfield-)Potts model as a statistical model for sequences $\underline{A} \in \mathcal{A}^L$, typically characterized by a fully connected coupling matrix J and local fields h (not represented). The model can be transformed into a Restricted Boltzmann Machine (RBM) by introducing Gaussian hidden variables $\underline{x} \in \mathbb{R}^p$, with p being the rank of J. Note the bipartite graphical structure of RBM, which causes the conditional probabilities $P(\underline{A}\,|\,\underline{x})$ and $P(\underline{x}\,|\,\underline{A})$ to factorize. **Panel B** shows a schematic representation of Persistent Contrastive Divergence (PCD). Initially the sample is initialized in the training data (the MSA of natural sequences), and then k alternating steps of sampling from $P(\underline{A}\,|\,\underline{x})$ resp. $P(\underline{x}\,|\,\underline{A})$ are performed. Parameters are updated after these k sampling steps, and sampling is continued using the updated parameters.

B. Parameter learning by persistent contrastive divergence

Maximizing the likelihood with respect to the parameters leads, for our RBM model, to the stationarity equations

$$\begin{aligned} \frac{1}{M}\sum_m \delta_{A_i^m,a} &= \langle \delta_{A_i,a}\rangle_{P(\underline{A},\underline{x})} \\ \frac{1}{M}\sum_m \delta_{A_i^m,a}\langle x^\mu\rangle_{P(\underline{x}\,|\,\underline{A}^m)} &= \langle \delta_{A_i,a} x^\mu\rangle_{P(\underline{A},\underline{x})} \end{aligned} \tag{21}$$

for all i, a and μ; the difference of both sides equals the gradient of the likelihood in direction of the corresponding parameter. While the first line matches the standard MaxEnt form – sample and ensemble average of an observable have to coincide, the second line contains a mixed sample-ensemble average on its left-hand side. Since the variables x^μ are latent and thus not contained in the MSA, an average over their probability $P(\underline{x}\,|\,\underline{A}^m)$ conditioned to the sequences $\underline{A}^m$ in the MSA has to be taken. Having a P-dependence on both sides of Eqs. (21) is yet another expression of the non-convexity of the likelihood function.

Model parameters $h_i(a)$ and $\xi_i^\mu(a)$ have to be fitted to satisfy the stationarity condition Eq. (21). This can be done iteratively: starting from arbitrarily initialized model parameters, we determine the difference between the left- and right-hand sides of this equation, and use this difference to update parameters (*i.e.* we perform gradient ascent of the likelihood); each of these update steps is called an *epoch* of learning. A major problem is that the exact calculation of averages over the $(L+p)$-dimensional probability distribution P is computationally infeasible. It is possible to estimate these averages by Markov Chain Monte Carlo (MCMC) sampling, but efficient implementations are needed since accurate parameter learning requires in practice thousands of epochs. To this aim, we exploit the bipartite structure of RBM: both conditional probabilities $P(\underline{A}\,|\,\underline{x})$ and $P(\underline{x}\,|\,\underline{A})$ are factorized. This allows us to initialise

MCMC runs in natural sequences from the MSA and to sample the $\underline{x}$ and the $\underline{A}$ in alternating fashion. As a second simplification we use *Persistent Contrastive Divergence* (PCD) [36]. Only in the first epoch the visible variables are initialised in the MSA sequences, and each epoch performs only a finite number of sampling steps (k for PCD-k), cf. Fig. 1.B. Trajectories are continued in a new epoch after parameter updates. If the resulting parameter changes become small enough, PCD will thereby generate close-to-equilibrium sequences, which form an (almost) *i.i.d.* sample of $P(\underline{A}, \underline{x})$ uncorrelated from the training set used for initialization.

Details of the algorithm, and comparison to the simpler contrastive divergence are given in the Appendix. Further technical details, like regularization, are also delegated to the Appendix.

C. Determining the likelihood contribution of single Hopfield-Potts patterns

It is obvious, that the total likelihood grows monotonously when increasing the number p of patterns ξ^μ. It is therefore important to develop criteria, which tell us if patterns are more or less important for modeling the protein family. To this aim we estimate the contribution of single patterns to the likelihood, by comparing the full model with a model, where a single pattern ξ^μ has been removed, while the other $p-1$ patterns and the local fields have been retained. The corresponding normalized change in log-likelihood reads

$$\Delta\ell_\mu = \frac{1}{M}\sum_{m=1}^{M}\left[\log P(\underline{A}^m) - \log P_{-\mu}(\underline{A}^m)\right] \ , \tag{22}$$

where $P_{-\mu}$ has the same form as given in Eq. (14) for P, but with pattern $\xi^\mu = \{\xi_i^\mu(a); i = 1, ..., L,\ a \in \mathcal{A}\}$ removed. Plugging Eq. (14) into Eq. (22) , we find

$$\Delta\ell_\mu = \frac{1}{M}\sum_{m=1}^{M}\left[\sum_{i=1}^{L}\xi_i^\mu(A_i^m)\right]^2 + \log\frac{Z_{-\mu}}{Z} \ . \tag{23}$$

The likelihood difference depends thus on the ratio of the two partition functions Z and $Z_{-\mu}$. While each of them is individually intractable due to the exponential sum over q^L sequences, the ratio can be estimated efficiently using importance sampling. We write:

$$\begin{aligned}\frac{Z_{-\mu}}{Z} &= \frac{1}{Z}\sum_{\underline{A}\in\mathcal{A}^L}\exp\left\{\sum_{\nu\neq\mu}\sum_{i,j=1}^{L}\xi_i^\nu(A_i)\xi_j^\nu(A_j) + \sum_{i=1}^{L}h_i(A_i)\right\} \\ &= \sum_{\underline{A}\in\mathcal{A}^L}P(\underline{A})\exp\left\{-\sum_{i,j=1}^{L}\xi_i^\mu(A_i)\xi_j^\mu(A_j)\right\} \ .\end{aligned} \tag{24}$$

The last expression contains the average of an exponential quantity over $P(\underline{A})$, so estimating the average by MCMC sampling of P might appear a risky idea. However, since P and $P_{-\mu}$ differ only in one of the p patterns, the distributions are expected to overlap strongly, and sufficiently large samples drawn from $P(\underline{A})$ can be used to estimate $Z_{-\mu}/Z$. Note that sampling is done from P, so the likelihood-contributions of all patterns can be estimated in parallel using a single large sample of the full model.

Once these likelihood contributions are estimated, we can sort them, and identify and interpret the patterns of largest importance in our Hopfield-Potts model.

V. HOPFIELD-POTTS MODELS OF PROTEIN FAMILIES

To understand the performance of Hopfield-Potts models in the case of protein families, we have analysed three protein families extracted from the Pfam database [2]: the Kunitz/Bovine pancreatic trypsin inhibitor domain (PF00014), the Response regulator receiver domain (PF00072) and the RNA recognition motif (PF00076). They have been selected since they have been used in DCA studies before, in our case RBM results will be compared to the ones of bmDCA, *i.e.* the generative version of DCA based on Boltzmann Machine Learning [29]. MSA are downloaded from the Pfam database [2], and sequences with more than 5 consecutive gaps are removed; cf. App. B for a discussion of the convergence problems of PCD-based inference in case of extended gap stretches. The resulting MSA dimensions for the three families are, in the order given before, $L = 52/112/70$ and $M = 10657/15000/10000$. As can be noted,

the last two MSA have been subsampled randomly since they were very large, and the running time of the PCD algorithm is linear in the sample size. The MSA for PF00072 was chosen to be slightly larger because of the longer sequences in this family.

In the following sections, results are described in detail for the PF00072 response regulator family. The results for the other protein families are coherent with the discussion; they are moved to the App. B for the seek of conciseness of our presentation.

A. Generative properties of Hopfield-Potts models

PCD is able, for all values of the pattern number p, to reach parameter values satisfying the stationarity conditions Eqs. (21). This is not only true when these are evaluated using the PCD sample propagated via learning from epoch to epoch, but also when the inferred model is resampled using MCMC, i.e. when the right-hand side of Eqs. (21) is evaluated using an *i.i.d.* sample of the RBM.

In the leftmost column of Fig. 2, this is shown for the single-site frequencies, i.e. for the first of Eqs. (21). The horizontal axis shows the statistics extracted from the original data collected in the MSA, while the vertical axis measures the same quantity in an *i.i.d.* sample extracted from the inferred model $P(\underline{A}, \underline{x})$. The fitting quality is comparable to the one obtained by bmDCA, as can be seen by comparison with the last panel in the first column of Fig. 2.

The other two columns of the figure concern the *generative* properties of RBM: connected two- and three-point correlations

$$\begin{aligned} c_{ij}(a,b) &= f_{ij}(a,b) - f_i(a)f_j(b) \\ c_{ijk}(a,b,c) &= f_{ijk}(a,b,c) - f_{ij}(a,b)f_k(c) - f_{ik}(a,c)f_j(b) - f_{jk}(b,c)f_i(a) + 2f_i(a)f_j(b)f_k(c) \end{aligned} \tag{25}$$

with the three-point frequencies $f_{ijk}(a,b,c)$ defined in analogy to Eqs. (6,9). Note that in difference to DCA, already the two-point correlations are not fitted directly by the RBM, but only the second moments related to the Hopfield-Potts patterns. This becomes immediately obvious for the case $p = 0$, where RBM reduce to simple profile models of statistically independent sites, but remains true for all values of $p < (q-1)L$. Note also that connected correlations are used, since the frequencies f_{ij} and f_{ijk} contain information about the fitted f_i, and therefore show stronger agreement between data and model.

The performance of RBM is found to be, up to statistical fluctuations, monotonous in the pattern number p. As in the mean-field approximation [31], no evident overfitting effects are observed. Even if not fitted explicitly, as few as $p = 20 - 40$ patterns are sufficient to faithfully reproduce even the non-fitted two- and three-point correlations. This is very astonishing, since only about 1.7-3.5% of the parameters of the full DCA model are used: the p patterns are given by $p(q-1)L$ parameters, while DCA has $(q-1)^2\binom{L}{2}$ independently inferred couplings. The times needed for accurate inference decrease accordingly: in some cases, a slight decrease in accuracy of bmDCA is observed as compared to RBM with the largest p; this could be overcome by iterating the inference procedure for further epochs.

B. Strong couplings and contact prediction

One of the main applications of DCA is the prediction of contacts between residues in the three-dimensional protein fold, based only on the statistics of homologous sequences. To this aim, we follow [25] and translate $q \times q$ coupling matrices $J_{ij}(a,b) = \sum_{\mu=1}^{p} \xi_i^\mu(a)\xi_j^\mu(b)$ for individual site pairs (i,j) into scalar numbers by first calculating their Frobenius norm,

$$F_{ij} = \sum_{a,b\in\mathcal{A}} J_{ij}(a,b)^2, \tag{26}$$

followed by the empirical average-product correction (APC)

$$F_{ij}^{APC} = F_{ij} - \frac{F_{i\cdot}F_{\cdot j}}{F_{\cdot\cdot}}\ , \tag{27}$$

where the $\cdot$ denotes an average over the corresponding index. The APC is intended to remove systematic non-functional bias due to conservation and phylogeny. These quantities are sorted, and the largest ones are expected to be contacts.

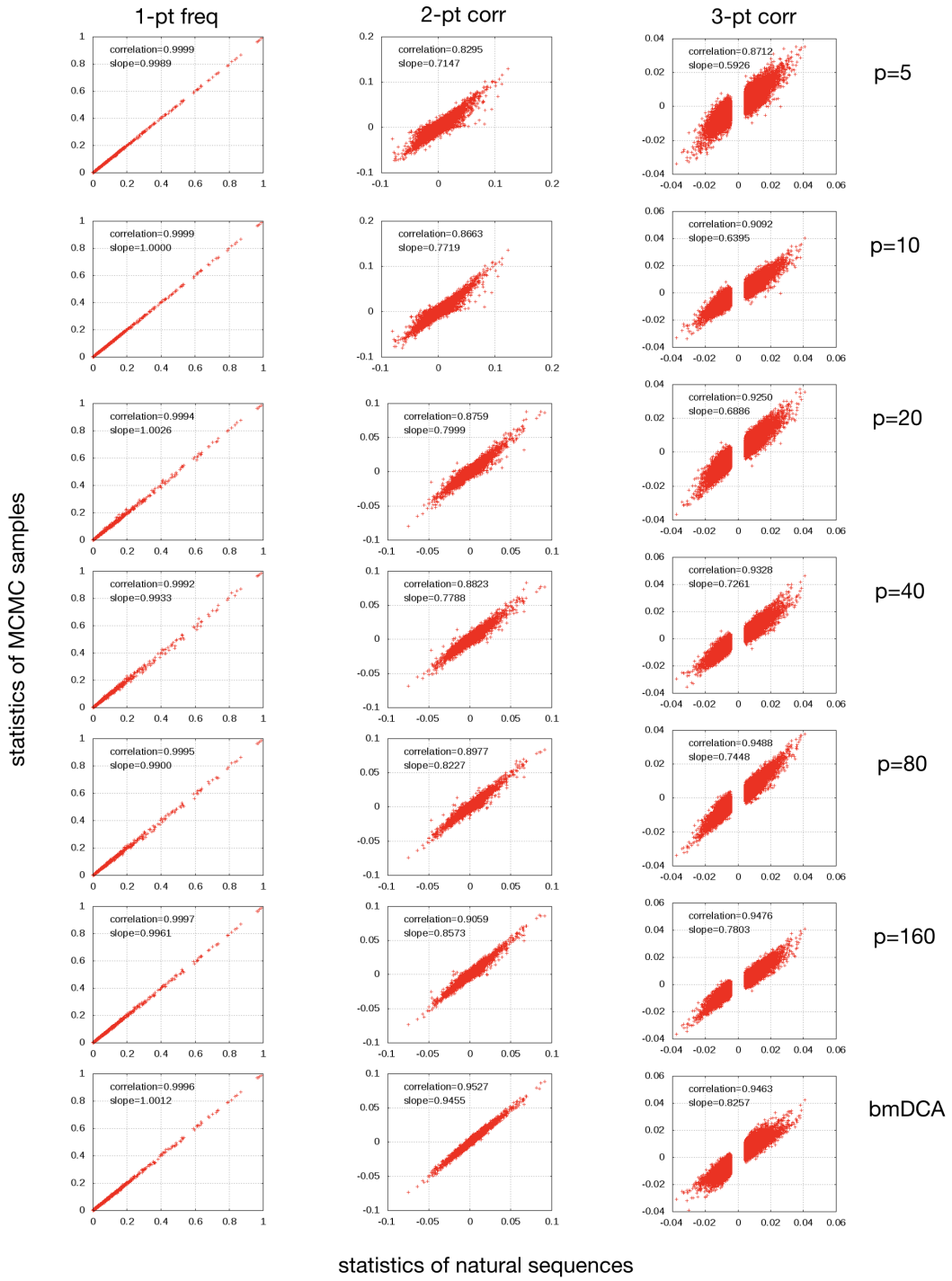

FIG. 2: Statistics of natural sequences (PF00072, horizontal axes) vs. MCMC samples (vertical axes) of Hopfield-Potts models for values of $p \in \{5, 10, 20, 40, 80, 160\}$ and for a full-rank Potts model inferred using bmDCA. The first column shows the 1-point frequencies $f_i(a)$ for all pairs (i, a) of sites and amino-acids, the other two column show the connected 2- and 3-point functions $c_{ij}(a, b)$ and $c_{ijk}(a, b, c)$. Due to the huge number of combinations for the three-point correlations, only the 100,000 largest values (evaluated in the training MSA) are shown. The Pearson correlations and the slope of the best linear fit are inserted in each of the panels.

The results for several values of p and for bmDCA are depicted in Fig. 3: the positive predictive value (PPV) is the fraction of true positives (TP) among the first n predictions, as a function of n. TP are defined as native contacts in a reference protein structure (PDB ID 3ilh [37] for PF00072), with a distance cutoff of 8Å. Pairs in vicinity along the

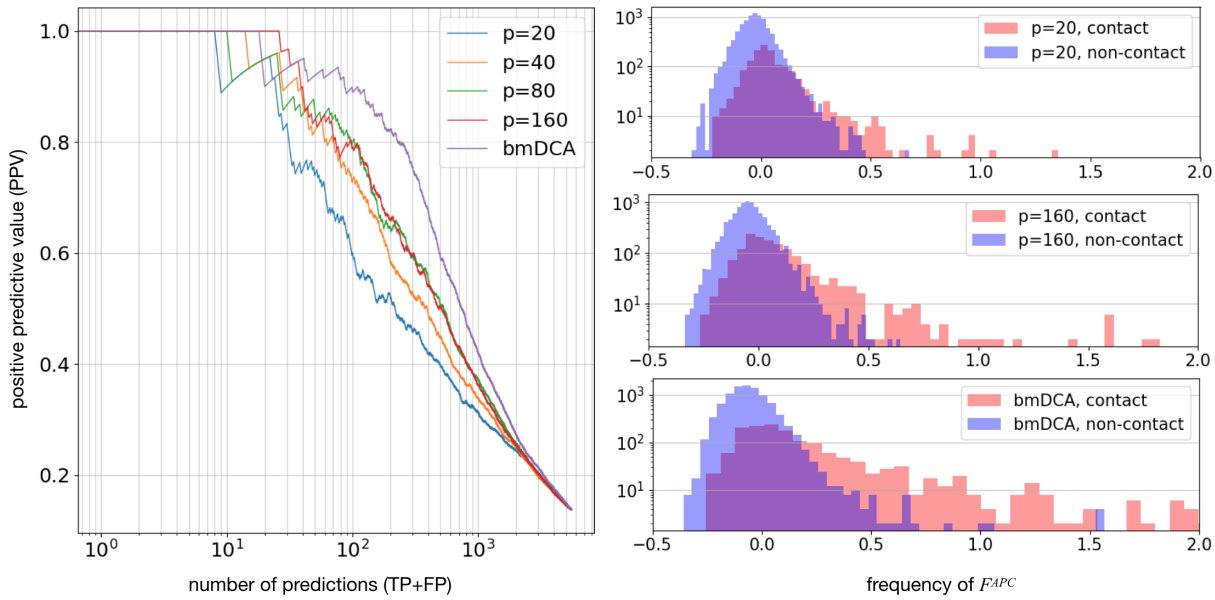

FIG. 3: The left panel shows the positive predictive value (PPV) for contact prediction as a function of the number of predictions, for various values of the pattern number p and for bmDCA. The left panels show, for $p = 20$, 160 and bmDCA, the distribution of coupling scores F_{ij}^{APC}. All residue pairs are grouped into contacts (red) and non-contacts (blue). The best contact predictions correspond to the positive tail of the red histogram, which becomes more pronounced when increasing p or even going to bmDCA.

peptide chain are not considered in this prediction, since they are trivially in contact: in coherence with the literature standard, Fig. 3 only considers predictions with $|i - j| \geq 5$.

Despite the fact, that even for as few as $p = 20 - 40$ patterns the model appears to be generative, *i.e.* non-fitted statistical observables are reproduced with good accuracy, the PPV curves depend strongly on the pattern number p. Up to statistically probably insignificant exceptions, we observe a monotonous dependence on p, and none of the RBM-related curves reaches the performance of the full-rank J_{ij} matrices of bmDCA. Even large values of p, where RBM have more than 30% of the parameters of the full Potts model, show a drop in performance in contact prediction.

Can we understand this apparent contradiction: similarly accurate reproduction of the statistics, but reduced performance in contact prediction? To this end, we consider in Fig. 3 the histogram of coupling strengths F_{ij}^{APC} divided into two subpopulations: values for sites i, j in contact are represented by red, for distant sites by blue histogram. It becomes evident that the rather compact histogram of non-contacts remains almost invariable with p (even is individual coupling values do change), but the histogram of contacts changes systematically: the tail of large F_{ij}^{APC} going beyond the upper edge of the blue histogram is less pronounced for small p. However, in the procedure described before, these F_{ij}^{APC}-values provide the first contact predictions.

The reduced capacity to detect contacts for small p is related to the properties of the Hopfield-Potts model in itself. While the strong signals form a sparse graph of residue-residue contacts, the Hopfield-Potts model is explicitly constructed to have a low-rank coupling matrix $(J_{ij}(a, b))$. It is, however, hard to represent a generic sparse matrix by a limited number of possibly distributed patterns. Hopfield-Potts models are more likely to detect distributed sequence signals than localized sparse ones.

C. Likelihood contribution and interpretation of selected sequence motifs

So what do the patterns represent? In Sec. IV C, we have discussed how to estimate the likelihood contribution of patterns, thereby being able to select the most important patterns in our model. Fig. 4 displays the ordered contributions for different values of p. We observe that, for small p, the distribution becomes more peaked, with few patterns having very large likelihood contributions. For larger p, the contributions are more distributed over many patterns, which collectively represent the statistical features of the data set.

Fig. 5 represents the first five patterns for $p = 20$. The left-hand site represents the pattern $\xi_i^\mu(a)$ as a sequence logo, a standard representation in sequence bioinformatics. Each site i corresponds to one position, the possible

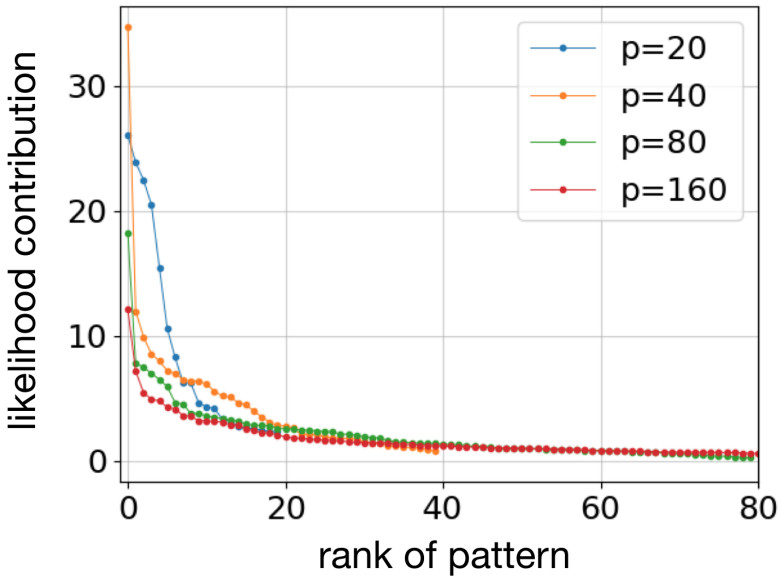

FIG. 4: Likelihood contribution of the individual patterns, for pattern numbers $p = 20, 40, 80, 160$.

amino-acids are shown by their one-letter code, the size of the letter being proportional to $|\xi_i^\mu(a)|$, according to the sign of $\xi_i^\mu(a)$, letters are represented above or below the zero line. The alignment gap is represented as a minus sign in an oval shape, which allows to represent its size in the current pattern.

Patterns are very distributed, both in terms of the sites and amino-acids with relatively large entries $\xi_i(a)$. This makes a direct interpretation of patterns complicated. It explains also why these patterns are not optimal to define localised contact predictions. Due to their distributed nature, they define groups of sites, which are connected via a dense network of comparable couplings. However, as we will see in the next section, the sites of large entries in a pattern define functional regions of proteins, which are important in sub-ensembles of proteins of strong (positive or negative) activity values along the pattern under consideration.

The middle column shows a histogram of pattern-specific activities of single sequences, *i.e.* of

$$x^\mu(\underline{A}) = \sum_{i=1}^{N} \xi_i^\mu(A_i) \ . \tag{28}$$

Note that, up to the rescaling in Eq. (16), these numbers coincide with the collective variables, or motifs, introduced in Eq. (11) at the beginning of this article. The blue histograms result from the natural sequences collected in the training MSA. They coincide well with the red histograms, which are calculated from an *i.i.d.* MCMC sample of our Hopfield-Potts model, including the bimodal structure of several histograms. This is quite remarkable: the Hopfield-Potts model was derived, in the beginning of this work, as the maximum-entropy model reproducing the first two moments of the activities $\{x^\mu(\underline{A}^m)\}_{m=1\ldots M}$. Finding higher-order features like bimodality is again an expression of the generative power of Hopfield-Potts models.

The right-most column of Fig. 5 proves the importance of individual patterns for the inferred model. The panels show the two-point correlations $c_{ij}(a,b)$ of the natural data (horizontal axis) vs. the one of samples drawn from the distributions $P_{-\mu}(\underline{A})$, introduced in Eq. (22) as Hopfield-Potts models of $p-1$ patterns, with pattern ξ^μ removed (vertical axis). The coherence of the correlations is strongly reduced when compared to the full model, which was shown in Fig. 2: removal even of a single pattern has a strong global impact on the model statistics.

D. Sequence clustering

As already mentioned, some patterns show a clear bimodal activity distribution, i.e. they identify two statistically distinct subgroups of sequences. The number of subgroups can be augmented by using more than one pattern, i.e. combinations of patterns can be used to cluster sequences.

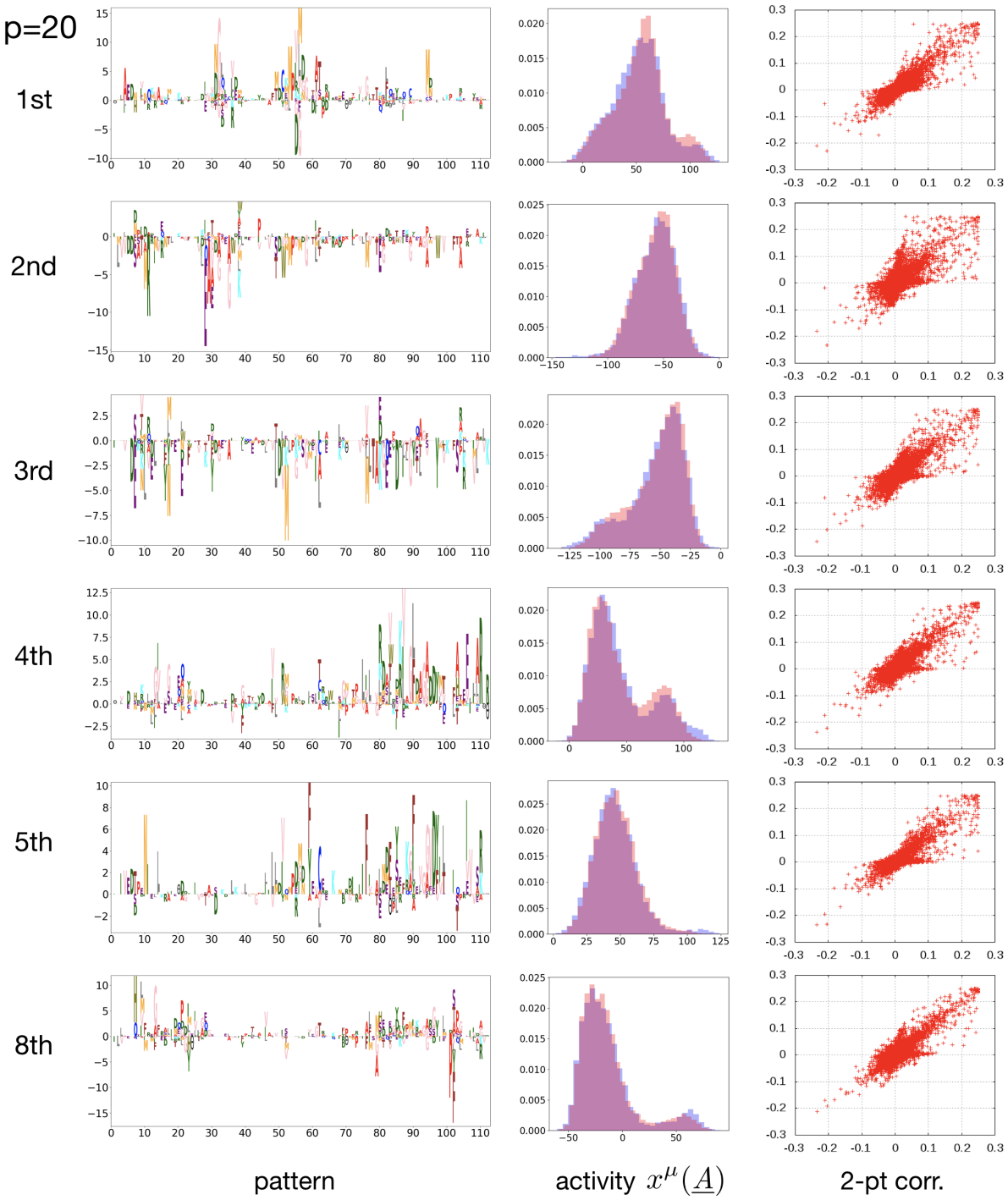

FIG. 5: The five patterns of highest likelihood-contribution for $p = 20$, the 8th ranking is added since used later in the text. The left panels show the patterns in logo representation, the letter-size is given by the corresponding element $\xi_i(a)$. The middle panel show the distribution of the activities, i.e. the projections of sequences onto the patterns. The blue histogram contains the natural sequences from the training MSA, the red histogram sequences sampled by MCMC from the Hopfield-Potts model. The right-hand site shows the connected 2-point correlations of the natural data (horizontal axis) vs. data sampled from $P_{-\mu}(\underline{A})$, i.e. a Hopfield-Potts model with one pattern removed. Strong deviations from the diagonal are evident.

To this aim, we have selected three patterns (number 6, 13 and 14) with a pronounced bimodal structure from the model with $p = 20$ patterns. In terms of likelihood contribution, they have ranks 8, 4 and 1 in the contributions to

the log-likelihood, cf. Fig. 5.

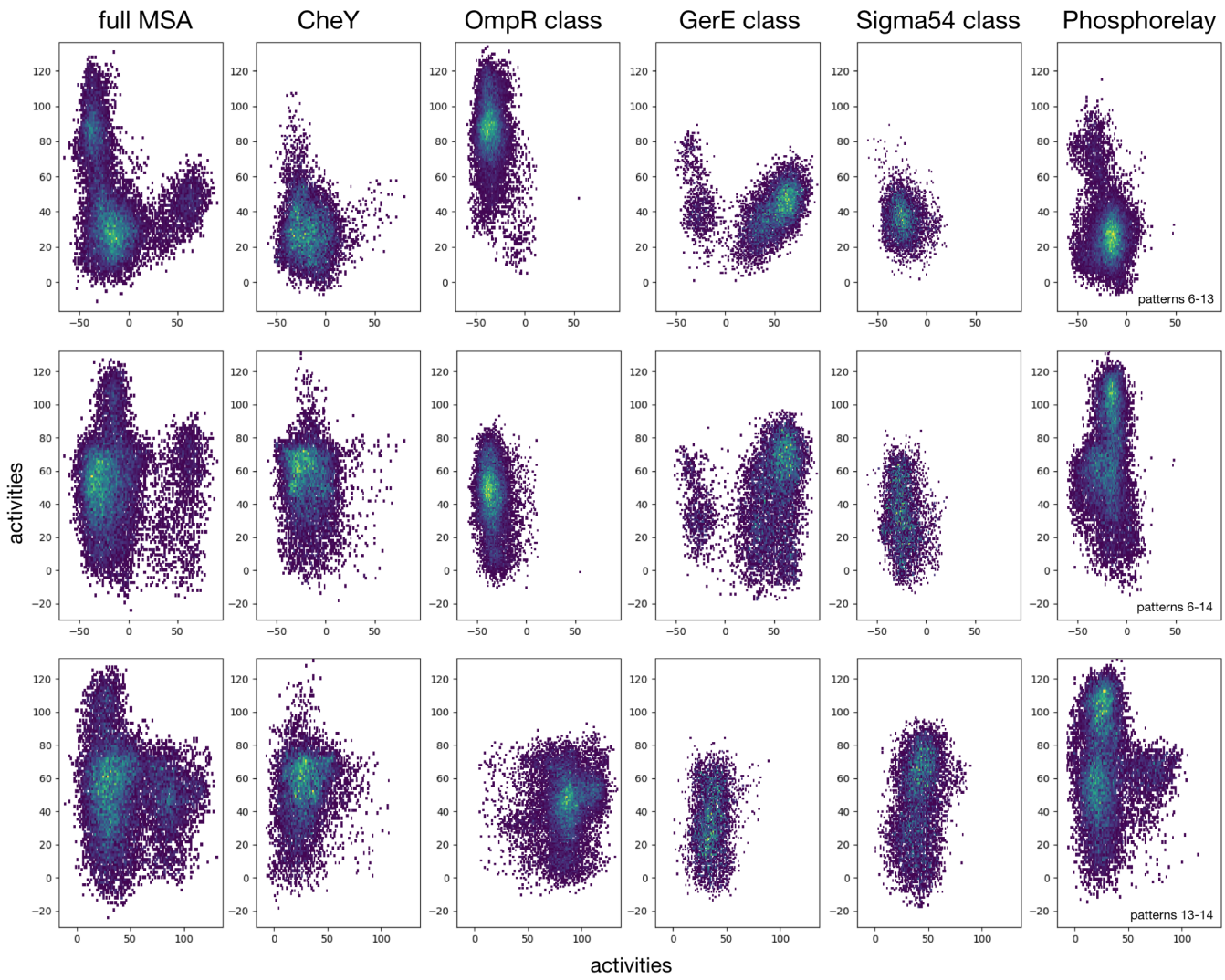

FIG. 6: Patterns with multimodal activity distributions for the set of all MSA sequences can be used to cluster sequences. The rows show, in this order, combinations of patterns 6-13, 6-14 and 13-14. Each sequence corresponds to a density-colored dot. A strongly clustered structure is clearly visible. When dividing the full MSA into functional subclasses, we can relate clusters to subclasses, and thus patterns to biological function.

The clustered organization of response-regulator sequences becomes even more evident in the two-dimensional plots characterizing simultaneously two activity distributions. The results for all pairs of the three patterns are displayed in Fig. 6, left column. As a first observation, we see that the main modes of the activity patterns give rise to one dominant cluster. Smaller cluster deviate from the dominant one in one single pattern, but show compatible activities in the other pattern - the two-dimensional plots therefore show typically an L-shaped sequence distribution, and three clusters, instead of the theoretically possible four combinations of activity models. It appears that single patterns identify the particularities of single subdominant sequence clusters.

We have chosen the response-regulator protein-domain family in this paper also due to the fact, that it constitutes a functionally well studied and diversified family. Response regulators are predominantly used in bacterial signaling systems:

- In *chemotaxis*, they appear as single-domain proteins named CheY, which transmit the signal from kinase proteins (activated by signal reception) to flagellar motor proteins, which trigger the movement of the bacteria. CheY proteins can be identified in our MSA as those coming from single-domain proteins, *i.e.* with lengths compatible to the PF00072-MSA width $L = 112$. We have selected a sub-MSA consisting of all proteins with total sequence lengths between 110 and 140 amino acids.

- In *two-component signal transduction* (TCS), response regulators are typically transcription factors, which are activated by signal-receiving histidine sensor kinases. The corresponding proteins contain two or three domains, in particular a DNA-binding domain, which is actually responsible for the transcription-factor activity of the activated response-regulator protein. According to the present DNA-binding domain, these TCS proteins can be subdivided into different classes, the dominant ones are the OmpR, the GerE and the Sigma54-HTH8 classes, we identified three sub-MSA corresponding to these classes by co-occurrence of the DNA-binding domains with the response-regulator domain in the same protein. The different DNA-binding domains are indicative for distinct homo-dimer structures assumed by the active transcription factors; DCA run on the sub-MS identifies these specific sub-family interfaces [38].

- *Phosphorelays* are similar to TCS, but consist of more complex multi-component signaling pathways. In these systems, found in bacteria and plants, response-regulator domains are typically fused to the histidine-sensor kinases. They do not act as transcription factors, but transduct a signal to a phosphotransferase, which finally activates a down-stream transcription factor of the same architecture mentioned in the last paragraph. We identified a class of response regulator domains, which are fused to a histidine kinase domain. In terms of domain architecture and protein length, this subfamily is extremely heterogenous.

Columns 2-6 in Fig. 6 show the activities of these five sub-families. It is evident, that distinct sub-MSA fall actually into distinct clusters according to these three patterns:

- The CheY-like single domain proteins (Col. 2) fall, according to all three patterns, into the dominant mode.

- The OmpR-class transcription factors (Col. 3) show a distinct distribution of higher activities for the second of the patterns (which actually has the most pronounced bimodal structure, probably due to the fact that the OmpR-class froms the largest sub-MSA). As can be seen in Fig. 6, this pattern has the largest positive entries in the region of positions 80-90 and 100-110. Interestingly, these regions define the interface of OmpR-class transcription-factor homodimerization, cf. [38]. In accordance with this structural interpretation, we also find a periodic structure of period 3-4 of the large entries in the pattern, which reflects the fact that the interface is formed by two helices, which lead to a periodic exposure of amino-acids in the protein surface.

- The GerE-class differs in activities in direction of the first pattern, only GerE-class proteins have positive, all other have negative activities. Dominant positive entries are found in regions 5-15 and 100-105, again identifying the homo-dimerization interface, cf. [38].

- The Sigma54 class does not show a distinct distribution of activities according to the three selected patterns. It is located together with the CheY-type sequences. However, when examining all patterns, we find that pattern number 5 (ranked 6th according to the likelihood contribution) is perfectly discriminating the two.

- Last but not least, the response-regulators fused to histidine kinases in phosphorelay systems show a distinct activity distribution according to the third pattern, mixing a part of activities compatible with the main cluster, and others being substantially larger (this mixing results presumably from the previously mentioned heterogenous structure of this sub MSA). Structurally known complexes between response-regulators and histidine phosphotransferases (PDB ID 4euk [39], 1bdj [40]) show the interface located in residues 5-15, 30-32 and 50-55, regions being important in the corresponding pattern. It appears that the pattern selects the particular amino-acid composition of this interface, which is specific to the phosphorelay sub-MSA.

These observations do not only show that the patterns allow for clustering sequences into subMSA, but the discriminating positions in the patterns have a clear biological interpretation. This is very intersting, since the analysis in [38] required an a-priori clustering of the initial MSA into sub-MSA, and the application of DCA to the individual sub-MSA. Here we have inferred only one Hopfield-Potts model describing the full MSA, and the patterns automatically identify biologically reasonable sub-families together with the sequence patterns characterizing them. The prior knowledge needed in [38] is not needed here; we use it only for the posterior interpretation of the patterns.

VI. CONCLUSION AND OUTLOOK

In this paper, we have rederived Hopfield-Potts models as statistical models for protein sequences with collective variable selection. The collective variables were assumed to have the form of additive sequence motifs, whose first and second moments are required to coincide between empirical data (the MSA of natural sequences) and the statistical sequence model. Within a maximum-entropy approach, these motifs (up to a rescaling factor) are found to be the Hopfield patterns defining the residue-residue couplings. In addition to the maximum-entropy framework, which is

built upon known observables, the Hopfield-Potts model adds a step of variable selection: the probability of the sequence data is maximised over all possible selections of motifs.

The quadratic coupling terms can be linearised using a Hubbard-Stratonovich transformation. When the Gaussian variables introduced in this transformation are interpreted as latent random variables, the Hopfield-Potts model takes the form of a Restricted Boltzmann Machine. This interpretation allows the application of efficient inference techniques, like persistent contrastive divergence, and therefore the accurate inference of the Hopfield-Potts patterns for any given MSA of a homologous protein family.

We find that Hopfield-Potts models acquire interesting generative properties even for a relatively small number of parameters (p=20-40). They are able to reproduce non-fitted properties like higher-order covariation of residues or features of the motif distribution, which cannot be encoded by its first two moments, e.g. bimodality. This bimodality in turn can be used to cluster sequences according to interpretable sequence motifs.

The Hopfield-Potts patterns, or sequence motifs, are typically found to be distributed over many residues, thereby representing global features of sequences. This observation explains, why Hopfield-Potts models tend to loose accuracy in residue-residue contact prediction, as compared to the full-rank Potts models used normally in Direct Coupling Analysis: the sparsity of the residue-residue contact network cannot be represented easily via few distributed sequence motifs, which describe more global patterns of sequence variability.

Individual sequences from the input MSA can be projected onto the Hopfield-Potts patterns, resulting in sequence-specific activity values. Some patterns show a mono-modal histogram for the protein family. They introduce a dense network of relatively small couplings between positions with sufficiently large entries in the pattern, without dividing the family into subfamilies. These patterns have great similarity to the concept of protein sectors, which was introduced in [41, 42] to detect distributed modes of sequence coevolution. Other patterns show a bimodal activity distributions, leading to the detection of functional sub-families. Since these are defined by, e.g., the positive vs. the negative entries of the pattern, the large entries in the pattern identify residues, which play a role similar to so-called specificity determining residues [43, 44], i.e. residues, which are conserved inside specific sub-families, but vary between sub-families. Both concepts emerge naturally in the context of Hopfield-Potts families.

These observations open up new ways of parameter reduction in statistical models of protein sequences: the sparsity of contacts, which are expected to be responsible for a large part of localized residue covariation in protein evolution, has to be combined with the low-rank structure of Hopfield-Potts models, which detect distributed functional sequence motifs. However, distributed patterns may also be related to phylogenetic correlations, which are present in the data, cf. [45]. As has been shown recently in a quite heuristic way [46], the decomposition of sequence-data covariance matrices or couplings matrices into a sum of a sparse and a low-rank matrix can substantially improve contact prediction, if only the sparse matrix is used.

Combining this idea with the idea of generative modeling seems a promising road towards parsimonious sequence models, which in turn would improve parameter interpretability and reduce overfitting effects, both limiting current versions of DCA.

Acknowledgements

We are particularly grateful to Pierre Barrat-Charlaix, Giancarlo Croce, Carlo Lucibello, Anna-Paola Muntoni, Andrea Pagnani, Edoardo Sarti and Francesco Zamponi for numerous discussions and assistance with data.

We acknowledge funding by the EU H2020 research and innovation programme MSCA-RISE-2016 under grant agreement No. 734439 InferNet. KS acknowledges an Erasmus Mundus TEAM (TEAM – Technologies for information and communication, Europe - East Asia Mobilities) scholarship of the European Commission in 2017/18, and a doctoral scholarship of the Honjo International Scholarship Foundation since 2018.

[1] UniProt Consortium, Nucleic Acids Research **47**, D506 (2018).
[2] R. D. Finn, P. Coggill, R. Y. Eberhardt, S. R. Eddy, J. Mistry, A. L. Mitchell, S. C. Potter, M. Punta, M. Qureshi, A. Sangrador-Vegas, et al., Nucleic acids research **44**, D279 (2016).
[3] S. R. Eddy, in *Genome Informatics 2009: Genome Informatics Series Vol. 23* (World Scientific, 2009), pp. 205–211.
[4] M. Remmert, A. Biegert, A. Hauser, and J. Söding, Nature Methods **9**, 173 (2012).
[5] T. Schwede, J. Kopp, N. Guex, and M. C. Peitsch, Nucleic Acids Research **31**, 3381 (2003).
[6] B. Webb and A. Sali, Current protocols in bioinformatics **47**, 5 (2014).
[7] D. De Juan, F. Pazos, and A. Valencia, Nature Reviews Genetics **14**, 249 (2013).
[8] S. Cocco, C. Feinauer, M. Figliuzzi, R. Monasson, and M. Weigt, Reports on Progress in Physics **81**, 032601 (2018).
[9] H. C. Nguyen, R. Zecchina, and J. Berg, Advances in Physics **66**, 197 (2017).

[10] M. Weigt, R. A. White, H. Szurmant, J. A. Hoch, and T. Hwa, Proceedings of the National Academy of Sciences **106**, 67 (2009).
[11] F. Morcos, A. Pagnani, B. Lunt, A. Bertolino, D. S. Marks, C. Sander, R. Zecchina, J. N. Onuchic, T. Hwa, and M. Weigt, Proceedings of the National Academy of Sciences **108**, E1293 (2011).
[12] S. Balakrishnan, H. Kamisetty, J. G. Carbonell, S.-I. Lee, and C. J. Langmead, Proteins: Structure, Function, and Bioinformatics **79**, 1061 (2011).
[13] D. T. Jones, D. W. Buchan, D. Cozzetto, and M. Pontil, Bioinformatics **28**, 184 (2012).
[14] D. S. Marks, L. J. Colwell, R. Sheridan, T. A. Hopf, A. Pagnani, R. Zecchina, and C. Sander, PloS one **6**, e28766 (2011).
[15] S. Ovchinnikov, H. Park, N. Varghese, P.-S. Huang, G. A. Pavlopoulos, D. E. Kim, H. Kamisetty, N. C. Kyrpides, and D. Baker, Science **355**, 294 (2017).
[16] J. K. Mann, J. P. Barton, A. L. Ferguson, S. Omarjee, B. D. Walker, A. Chakraborty, and T. Ndung'u, PLoS Comput Biol **10**, e1003776 (2014).
[17] F. Morcos, N. P. Schafer, R. R. Cheng, J. N. Onuchic, and P. G. Wolynes, Proceedings of the National Academy of Sciences **111**, 12408 (2014).
[18] M. Figliuzzi, H. Jacquier, A. Schug, O. Tenaillon, and M. Weigt, Molecular Biology and Evolution **33 (1)**, 268 (2016).
[19] H. Szurmant and M. Weigt, Current Opinion in Structural Biology **50**, 26 (2018).
[20] J. Tubiana, S. Cocco, and R. Monasson, eLife **8**, e39397 (2019).
[21] E. T. Jaynes, Physical review **106**, 620 (1957).
[22] R. Durbin, S. R. Eddy, A. Krogh, and G. Mitchison, *Biological sequence analysis: probabilistic models of proteins and nucleic acids* (Cambridge university press, 1998).
[23] S. R. Eddy, Bioinformatics **14**, 755 (1998).
[24] C. Baldassi, M. Zamparo, C. Feinauer, A. Procaccini, R. Zecchina, M. Weigt, and A. Pagnani, PloS ONE **9**, e92721 (2014).
[25] M. Ekeberg, C. Lövkvist, Y. Lan, M. Weigt, and E. Aurell, Physical Review E **87**, 012707 (2013).
[26] L. Sutto, S. Marsili, A. Valencia, and F. L. Gervasio, Proceedings of the National Academy of Sciences **112**, 13567 (2015).
[27] A. Haldane, W. F. Flynn, P. He, R. Vijayan, and R. M. Levy, Protein Science **25**, 13781384 (2016).
[28] J. P. Barton, E. De Leonardis, A. Coucke, and S. Cocco, Bioinformatics **32**, 3089 (2016).
[29] M. Figliuzzi, P. Barrat-Charlaix, and M. Weigt, Molecular Biology and Evolution **35**, 1018 (2018).
[30] M. Socolich, S. W. Lockless, W. P. Russ, H. Lee, K. H. Gardner, and R. Ranganathan, Nature **437**, 512 (2005).
[31] S. Cocco, R. Monasson, and M. Weigt, PLoS computational biology **9**, e1003176 (2013).
[32] G. D. Stormo, T. D. Schneider, L. Gold, and A. Ehrenfeucht, Nucleic Acids Research **10**, 2997 (1982).
[33] E. van Nimwegen, BMC Bioinformatics **8**, S4 (2007).
[34] P. Smolensky, Tech. Rep., Colorado Univ at Boulder Dept of Computer Science (1986).
[35] G. E. Hinton and R. R. Salakhutdinov, science **313**, 504 (2006).
[36] G. E. Hinton, in *Neural networks: Tricks of the trade* (Springer, 2012), pp. 599–619.
[37] B. Syed Ibrahim, S. K. Burley, and S. Swaminathan (2009), released on PDB by: New York SGX Research Center for Structural Genomics (NYSGXRC).
[38] G. Uguzzoni, S. J. Lovis, F. Oteri, A. Schug, H. Szurmant, and M. Weigt, Proceedings of the National Academy of Sciences **114**, E2662 (2017).
[39] J. Bauer, K. Reiss, M. Veerabagu, M. Heunemann, K. Harter, and T. Stehle, Molecular Plant **6**, 959 (2013).
[40] M. Kato, T. Shimizu, T. Mizuno, and T. Hakoshima, Acta Crystallographica Section D: Biological Crystallography **55**, 1257 (1999).
[41] N. Halabi, O. Rivoire, S. Leibler, and R. Ranganathan, Cell **138**, 774 (2009).
[42] O. Rivoire, K. A. Reynolds, and R. Ranganathan, PLoS Computational Biology **12**, e1004817 (2016).
[43] G. Casari, C. Sander, and A. Valencia, Nature structural biology **2**, 171 (1995).
[44] L. A. Mirny and M. S. Gelfand, Journal of molecular biology **321**, 7 (2002).
[45] C. Qin and L. J. Colwell, Proceedings of the National Academy of Sciences **115**, 690 (2018).
[46] H. Zhang, Y. Gao, M. Deng, C. Wang, J. Zhu, S. C. Li, W.-M. Zheng, and D. Bu, Biochemical and Biophysical Research Communications **472**, 217 (2016).
[47] A. Wlodawer, J. Walter, R. Huber, and L. Sjölin, Journal of Molecular Biology **180**, 301 (1984).
[48] C. Pancevac, D. C. Goldstone, A. Ramos, and I. A. Taylor, Nucleic Acids Research **38**, 3119 (2010).

Appendix A: Results for other protein families

The first appendix is dedicated to other protein families. As discussed in the main text, we have analyzed three distinct families, and discussed only one in full detail in the main text. Here we present the major results – generative properties, contact prediction and selected collective variables (patterns) – for two more families. These results show the general applicability of our approach beyond the specific response-regulator family used in the main text.

1. Kunitz/Bovine pancreatic trypsin inhibitor domain PF00014

Figs. 7, 8 et 9 display the major results for the PF00014 protein family. PPV curves are calculated using PDB ID 5pti [47].

2. RNA recognition motif PF00076

Figs. 10, 11 et 12 display the major results for the PF00076 protein family. PPV curves are calculated using PDB ID 2x1a [48].

Appendix B: Notes and details on inference methods

1. Regularization

In case of limited data but many parameters, i.e. the case (Hopfield-)Potts models for protein families are in, the direct likelihood maximisation in Eq. (18) can lead to overfitting effects, causing problems in sampling and parameter interpretation. To give a simple example, a rare and therefore unobserved event would be assigned zero probability, corresponding to (negative) infinite parameter values.

To cope with this problem, regularization is used. Regularization in general penalizes large (resp. non-zero) parameter values, and can be justified in Bayesian inference as a prior distribution acting on the parameter values. In this paper and following [20], we use a block regularization of the form

$$R(\xi, h) = \eta_0 \sum_{\mu=1}^{p} \left(\sum_{i,a} |\xi_i^\mu(a)| \right)^2 + q\eta_0 \sum_{i,a} h_i(a)^2 \ , \tag{B1}$$

with η_0 being a hyperparameter determining the strength of regularization. This regularization weakly favors sparsity of the patterns.

We use $\eta_0 = \alpha_0 L/qM$ with $\alpha_0 = 0.0525$ as default values throughout this paper. In the last section of this appendix, we show that Hopfield-Potts inference is robust with respect to this choice.

2. Contrastive divergence vs. persistent contrastive divergence

a. *Contrastive divergence does not reproduce the two-point statistics*

Contrastive divergence (CD) is a method for training restricted Boltzmann machines similar to persistent contrastive divergence. Initialized in the original data, i.e. the MSA of natural amino-acid sequences, a few sampling steps are performed in analogy to Fig. 1, the kth step is used in the parameter update to approach a solution of Eq. (21). However, rather than continuing the MCMC sampling from this sample, the sample is re-initialized in the original data after each epoch. This has, a priori, advantages and disadvantages: The sample remains close to a good sample of the model in CD, but far from a sample of the intermediate model with not yet converged parameters.

As can be seen in Fig. 13, after a sufficient number of epochs the statistics of the CD sample and the training data are perfectly coherent, the model appears to be converged. However, the connected two-point correlations are not well reproduced when resampling the inferred model with standard MCMC. Part of the empirically non-zero correlations are not reproduced and mistakenly assigned very small values in the inferred model.

To understand this observation, we have selected the elements of the second panel, which show discrepancies between empirical and model statistics, cf. the insert in the figure. The corresponding values of (i, j, a, b) are strongly localized in the beginning and the end of the protein chain, and correspond to the gap-gap statistics $c_{ij}(-,-)$. This gives a strong hint towards the origin of the problems in CD-based model inference: gap stretches, which exist in MSA of natural sequences in particular at the beginning and the end of proteins, due to the local nature of the alignment algorithm used in Pfam. Those located at the beginning of the sequence start in position 1, and continue with only gaps until they are terminated by an amino-acid symbol. They never start later than in position 1 or include individual internal amino-acid symbols (analogous for the gap stretches at the end, which go up to the last position $i = L$).

In CD only a few sampling steps are performed, so stretched gaps in the initialization tend to be preserved even if the associated gap-gap couplings are very weak. Basically to remove a gap stretch, an internal position can not be

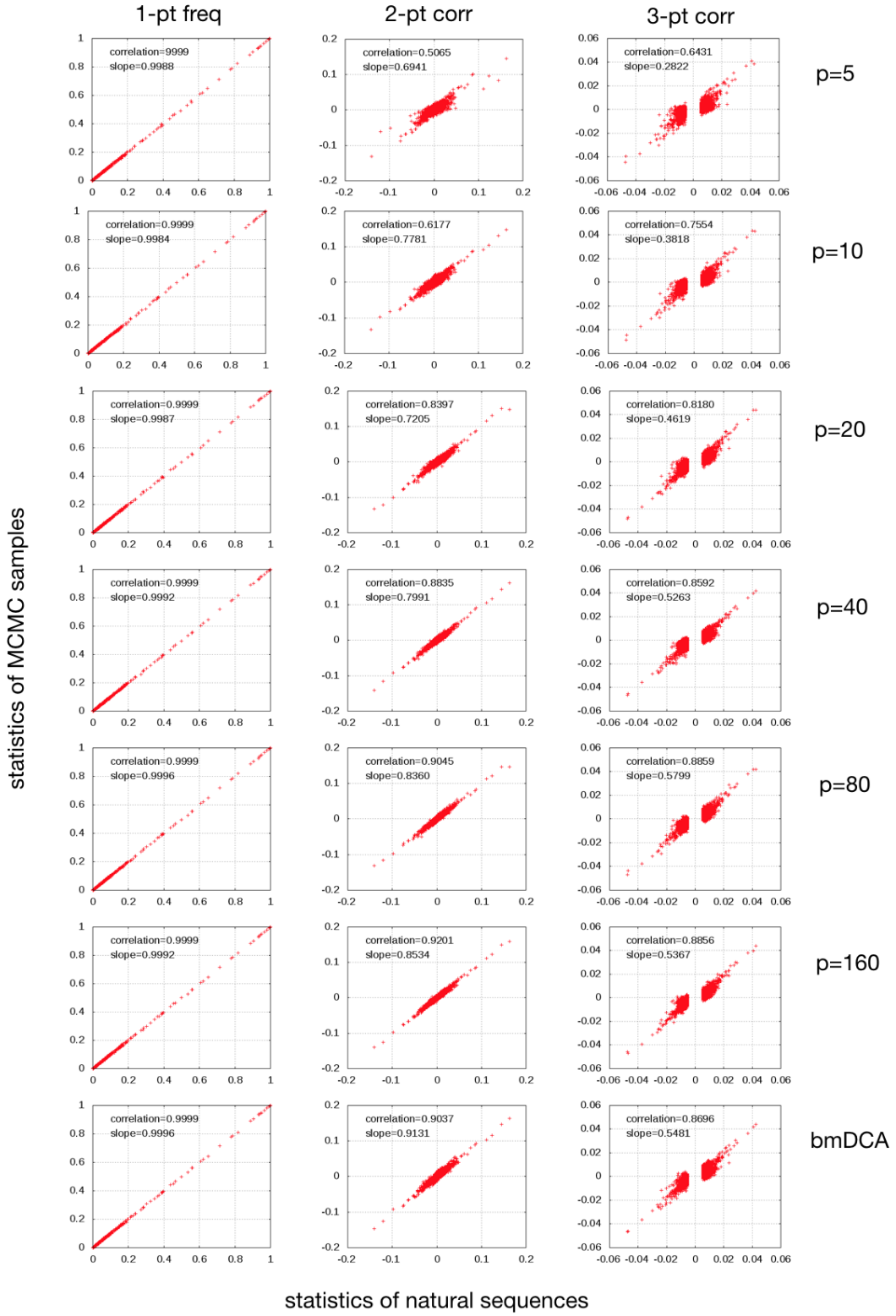

FIG. 7: Same as Fig. 2, but for the protein family PF00014.

switched to an amino-acid, but the gap has to be removed iteratively from one of ts endpoints, namely the one inside the sequence (i.e. not positions 1 or L). So, in CD, even small couplings are thus sufficient to reproduce the gap-gap statistics.

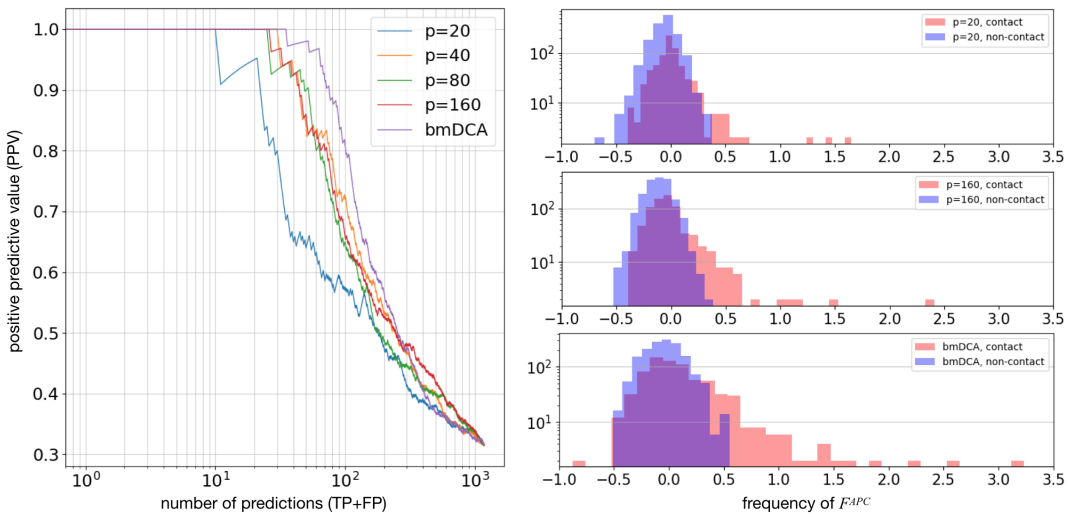

FIG. 8: Same as Fig. 3, but for the protein family PF00014.

If resampling the same model with MCMC, parameters have to be such that gap-stretches emerge spontaneously during sampling. This requires quite large couplings, actually in bmDCA gap-gap couplings between neighboring sites are the largest couplings of the entire Potts model. Using now the small couplings inferred by CD, these gap-stretches do not emerge at sufficient frequency, and correspondingly the positions at the extremities of the sampled sequences appear less correlated.

b. *Persistent contrastive divergence and transient oscillations in the two-point statistics*

Persistent contrastive divergence overcomes this sampling issue by not reinitialising the sample after each epoch, but by continuing the MCMC exploration in the next epoch, with updated parameters.

As shown in the main text, PCD can actually be used to infer parameters, which lead to accurately reproduced two-point correlations when *i.i.d.* samples are generated from the inferred model. However, during inference we have observed transient oscillations, cf. Fig. 14 for an epoch where correlations in a subset of positions and amino acids are overestimated. An analysis of the of the positions and the spins involved in these deviations shows, that again gap stretches are responsible.

The reason can be understood easily. Initially PCD is not very different from CD. Gap stretches are present due to the correlation with the training sample, and only small gap-gap interactions are learned. However, after a some epochs the sample will loose the correlation with the training sample. Due to the currently small gap-gap correlations, gap stretches are lost in the PCD sample. According to our update rules, the corresponding gap-gap couplings will fastly increase. However, due to the few sampling steps performed in each PCD epoch, this growth will go on even when parameters would be large enough to generate gap stretches in an *i.i.d.* sample. Also in the PCD sample, gap stretches will now emerge, but due to the overestimation of parameters, they will be more frequent than in the training sample, i.e. parameters start to decrease again. An oscillation of gap-gap couplings is induced.

The strength of these oscillations can be strongly reduced by removing samples with large gap stretches from the training data, and train only on data with limited gaps. If the initial training set was large enough, the resulting models are even expected to be more precise, since gap stretches do contain no or little information about the amino-acid sequences under study. However, if samples are too small, the suggested pruning procedure may reduce the sample to an insufficient size for accurate inference. Care has thus to be taken when removing sequences.

3. Robustness of the results

As discussed before, we need to include regularization to avoid overfitting due to limited data. In Figs. 15 and 16, we show the dependence of the inference results due to changes of the regularization strength over roughly two orders of magnitude. The first of the two figures shows the results for CD: empirical connected two-point correlation

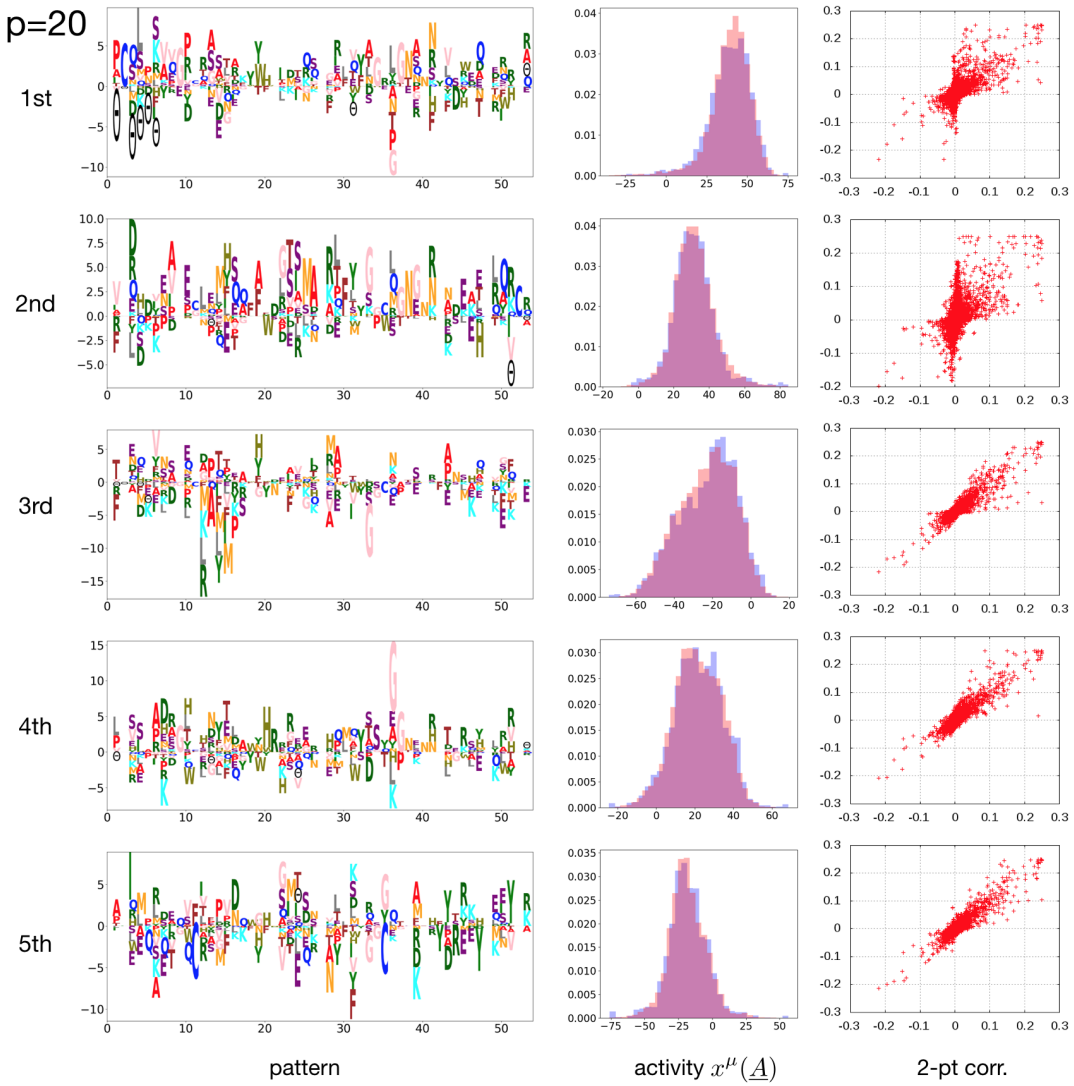

FIG. 9: Same as Fig. 5, but for the protein family PF00014.

are compared with *i.i.d.* samples of the corresponding models. We note that the results depend strongly on the regularization strength. For low regularization, the correspondence between model and MSA is low, due to overfitting. At strong regularization, only part of the correlations is reproduced, we over-regularize and thus underfit the data. For each protein family, and each number p of patterns, the regularization strength would have to be tuned.

For PCD, the situation is fortunately much better, results are found to be very robust with respect to regularization, cf. Fig. 16. This allows us to choose one regularization strength across protein families and pattern numbers.

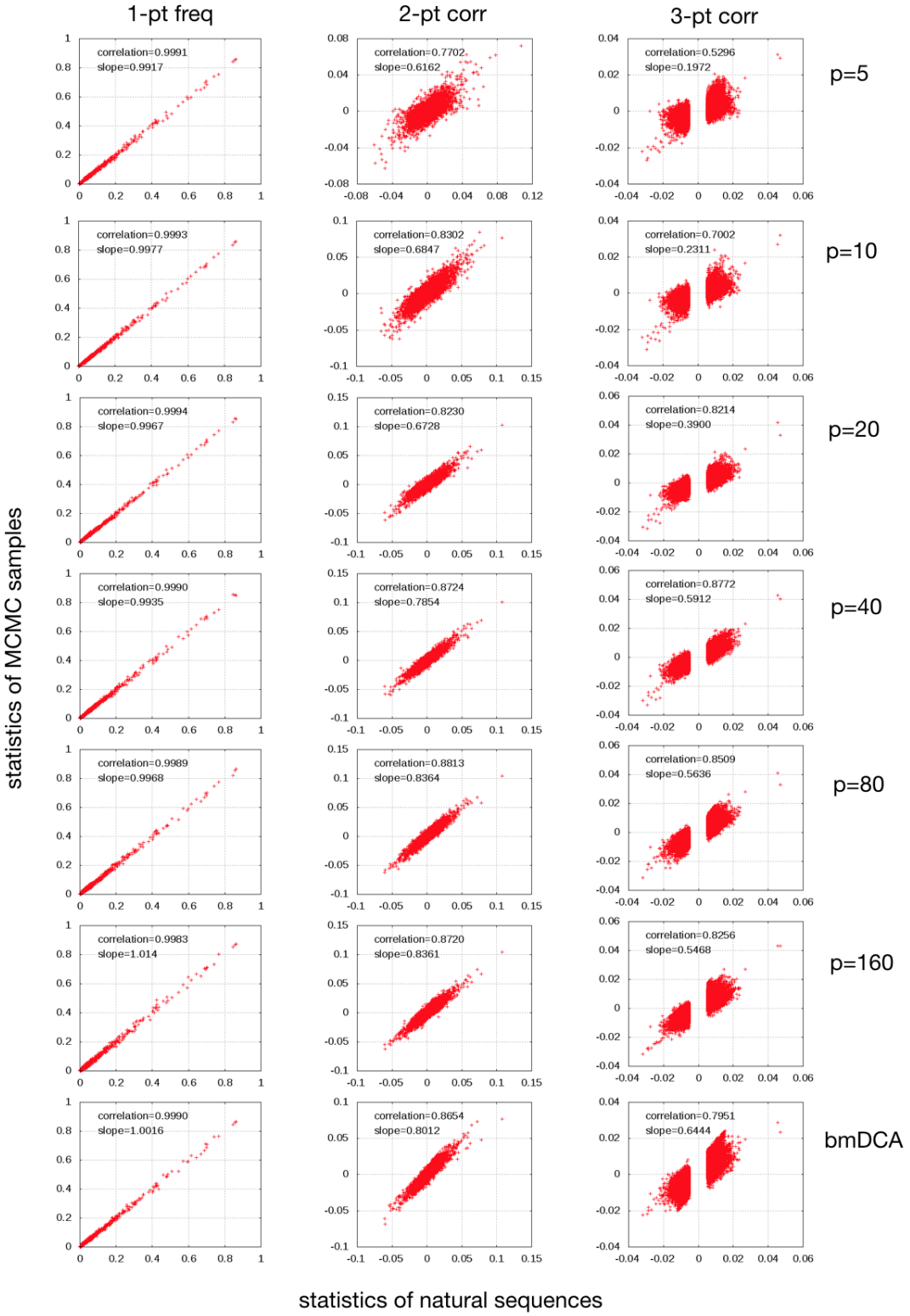

FIG. 10: Same as Fig. 2, but for the protein family PF00076.

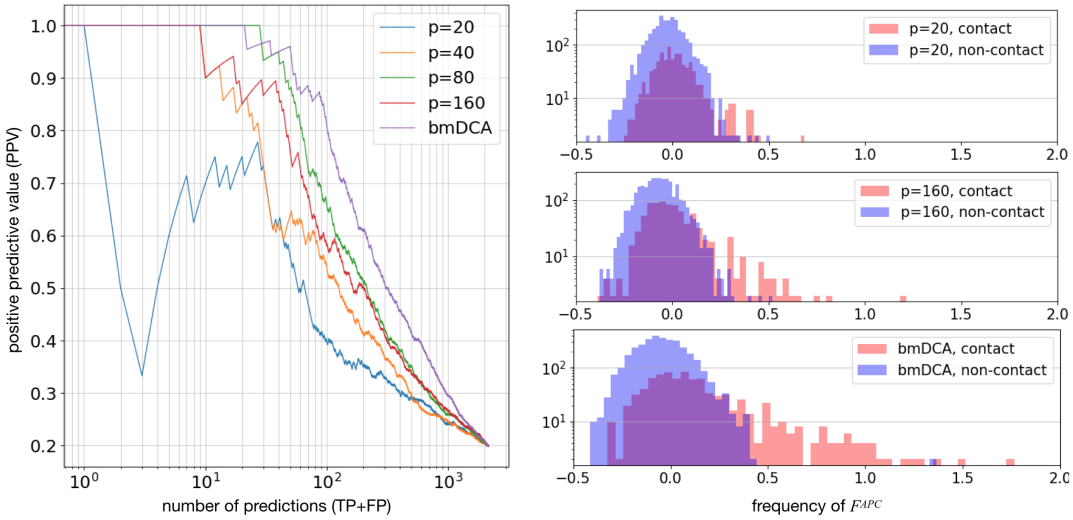

FIG. 11: Same as Fig. 3, but for the protein family PF00076.

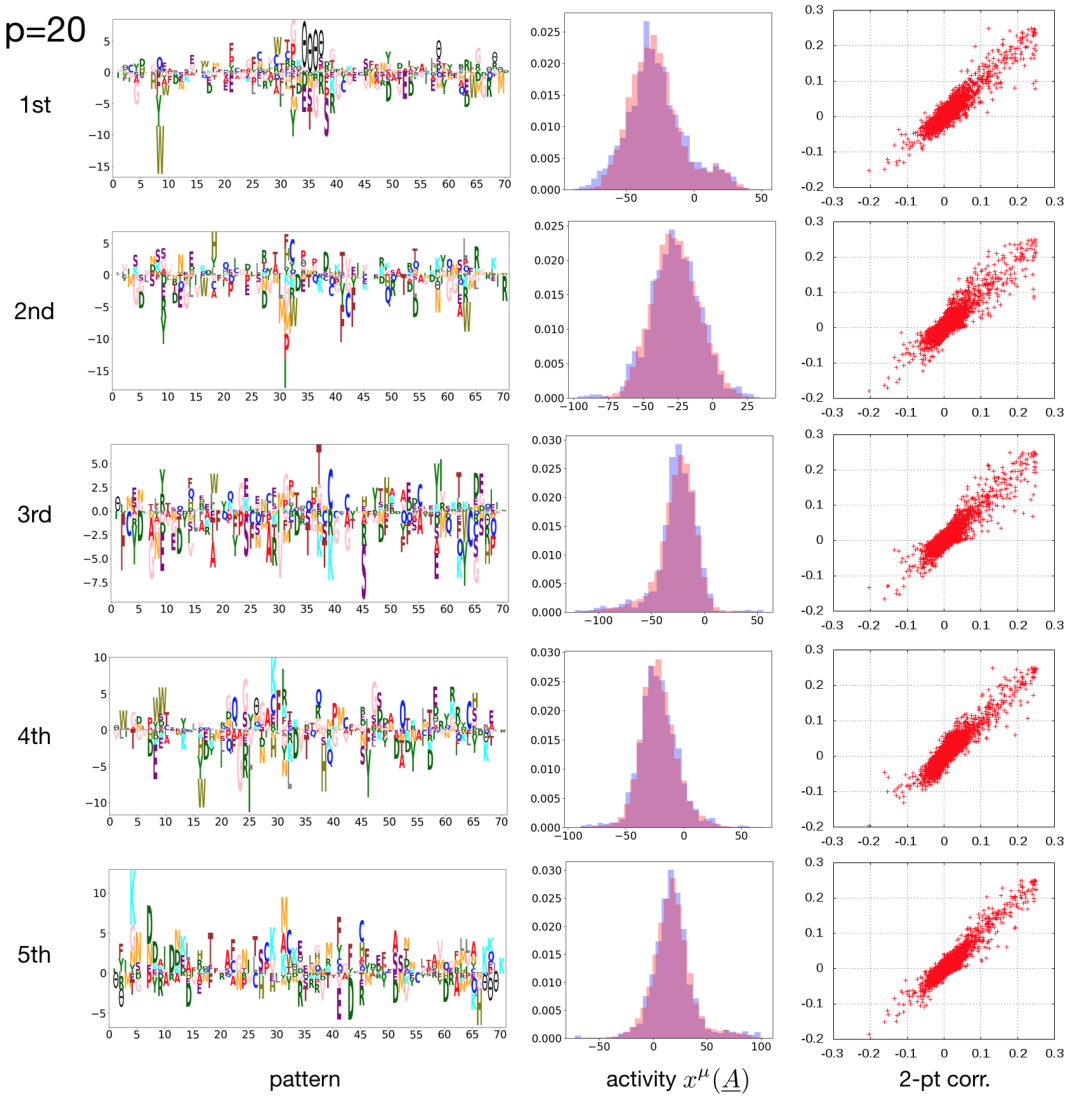

FIG. 12: Same as Fig. 5, but for the protein family PF00076.

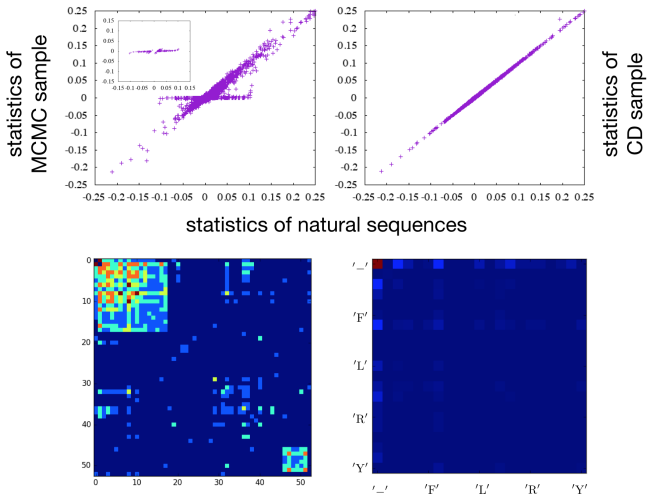

FIG. 13: The upper two panels show the statistics (two-point connected correlations) of the training data (PF00014) against an *i.i.d.* MCMC sample extracted from the inferred model, and the CD sample used for inference. The perfect coincidence of the two in the CD case demonstrates that the CD algorithm is converged, contrast is lost despite the fact, that the correlations extracted from an *i.i.d.* sample do not match the empirical ones. To understand this, we have selected those (i, j, a, b) with substantial deviations, cf. the insert in the first panel, and analyzed their location in the protein (first panel) and their amino-acid composition (second panel, amino-acids in alphabetical order of one letter code $[-, A, C, ...Y]$), densities are represented via heatmap plots. Location at the extremities od the sequences and in gap-gap correlations emerge clearly.

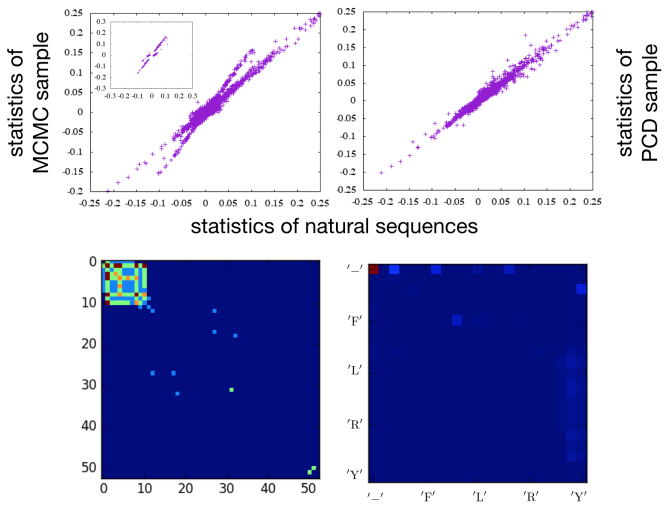

FIG. 14: The upper two panels show the statistics (two-point connected correlations) of the training data (PF00014) against an *i.i.d.* MCMC sample extracted from the inferred model, and the PCD sample used for inference. As in the CD case, the MCMC sample shows larger deviations from the empirical observations than the PCD sample, but correlations appear overestimated, and contrast in the PCD plot is sufficient to drive further evolution of parameters. To understand these observation, we have selected those (i, j, a, b) with substantial deviations, cf. the insert in the first panel, and analyzed their location in the protein (first panel) and their amino-acid composition (second panel, amino-acids in alphabetical order of one letter code $[-, A, C, ...Y]$), densities are represented via heatmap plots. Location at the extremities od the sequences and in gap-gap correlations emerge clearly.

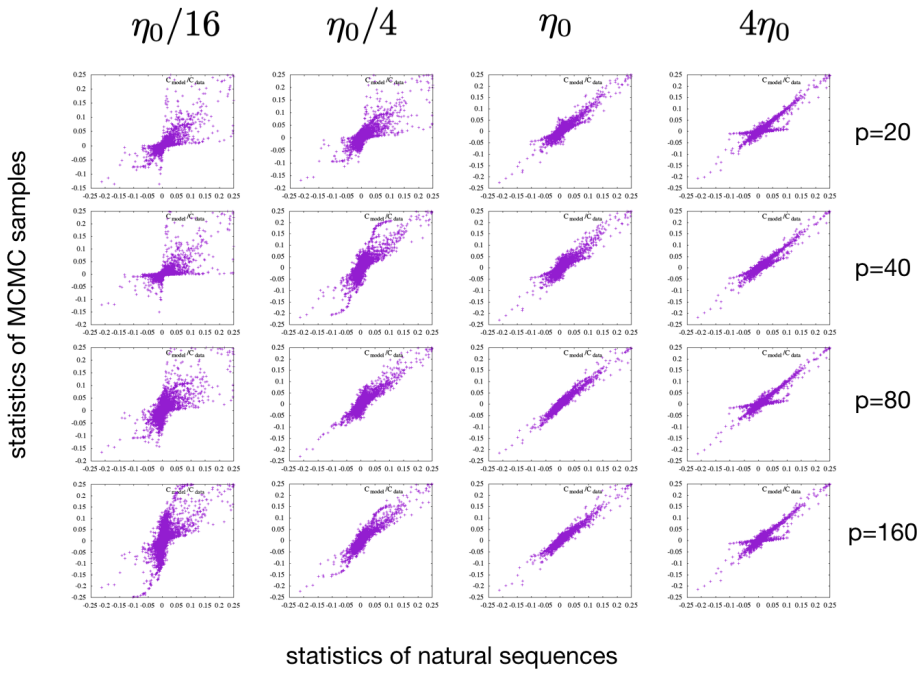

FIG. 15: Regularization dependence for CD inference, empirical two-point connected correlations (PF00014) are plotted against those estimated from the model using an *i.i.d.* MCMC sample. The regularization strength is varied over almost two orders of magnitude, with $\eta_0 = \alpha_0 L/qM$ and $\alpha_0 = 0.0525$, going from a zone of overfitting to one of over-regularization. Results are shown for various values of p, illustrating a strong p-dependence of the optimal coupling strength.

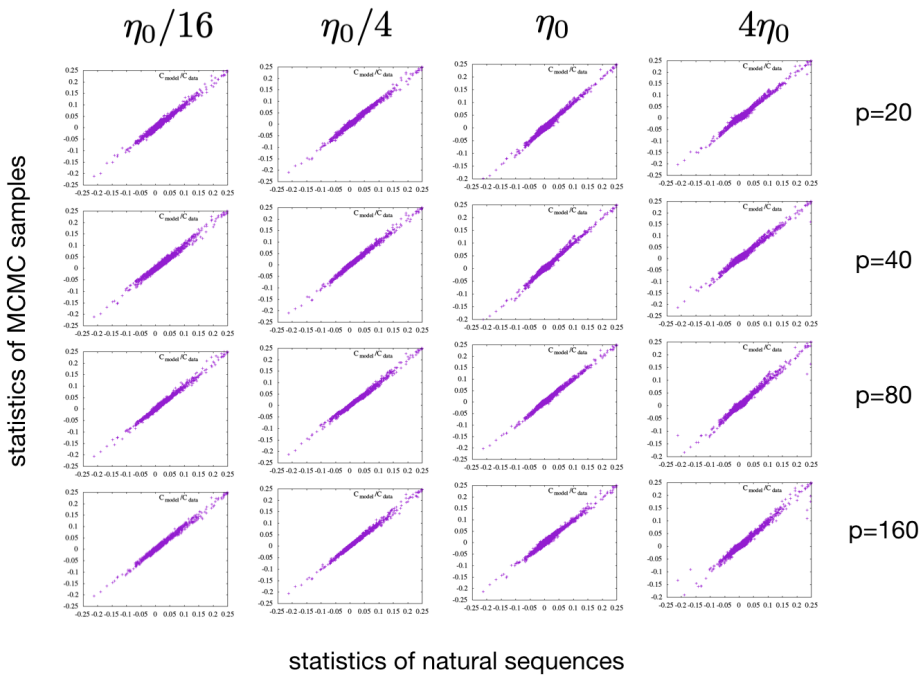

FIG. 16: Regularization dependence for PCD inference, empirical two-point connected correlations (PF00014) are plotted against those estimated from the model using an *i.i.d.* MCMC sample. The regularization strength is varied over almost two orders of magnitude, with $\eta_0 = \alpha_0 L/qM$ and $\alpha_0 = 0.0525$; results are shown for various values of p. We find a strong robustness of results with respect to regularization.