Random versus maximum entropy models of neural population activity

Ulisse Ferrari,[1] Tomoyuki Obuchi,[2] and Thierry Mora[3,*]
[1]*Institut de la Vision, INSERM and UMPC, 75012 Paris, France*
[2]*Department of Mathematical and Computing Science, Tokyo Institute of Technology, Yokohama 226-8502, Japan*
[3]*Laboratoire de physique statistique, École normale supérieure, CNRS and UPMC, 75005 Paris, France*
(Received 9 December 2016; published 27 April 2017)

The principle of maximum entropy provides a useful method for inferring statistical mechanics models from observations in correlated systems, and is widely used in a variety of fields where accurate data are available. While the assumptions underlying maximum entropy are intuitive and appealing, its adequacy for describing complex empirical data has been little studied in comparison to alternative approaches. Here, data from the collective spiking activity of retinal neurons is reanalyzed. The accuracy of the maximum entropy distribution constrained by mean firing rates and pairwise correlations is compared to a random ensemble of distributions constrained by the same observables. For most of the tested networks, maximum entropy approximates the true distribution better than the typical or mean distribution from that ensemble. This advantage improves with population size, with groups as small as eight being almost always better described by maximum entropy. Failure of maximum entropy to outperform random models is found to be associated with strong correlations in the population.

DOI: 10.1103/PhysRevE.95.042321

I. INTRODUCTION

The principle of maximum entropy was introduced in 1957 by Jaynes [1,2] to formulate the foundations of statistical mechanics as an inference problem. Its interest has been recently rekindled by its application to a variety of data-rich fields, starting with the correlated activity of populations of retinal neurons [3,4]. The method has since been used to study correlations in other neural data, such as cortical networks [5–7] and functional magnetic resonance imaging [8], as well as in other biological and nonbiological contexts, including multiple sequence alignments of proteins [9–11] and nucleic acids [12,13], the collective motion of bird flocks [14], the spelling rules of words [15], and the statistics of decisions by the United States Supreme Court [16]. In many cases, the close link between maximum entropy and statistical mechanics has led to new insights into the thermodynamics of the system in terms of phase transitions [17–19], or multivalley energy landscape [20,21]. In other cases, the method has allowed for predictions of crucial practical relevance, such as residue contacts in proteins [22], or deleterious mutations in HIV [23].

Although the motivations of maximum entropy seem intuitive and can be formalized rigorously [24], the perceived arbitrariness of its assumptions has led to question its validity [25,26]. The starting point is to consider models that match empirical observations on a few key statistics of the data. Maximum entropy's crucial—and debatable—assumption is to pick, out of the many models that satisfy that constraint, the one with the largest Gibbs entropy. This choice seems natural, since it ensures that the model is as random as possible. However, it is not clear why it should describe the data better than other models satisfying the same constraints. To address this question directly on empirical data, we reanalyze the original neural data from Ref. [3], which contributed to the recent surge of interest in maximum entropy. We compare the accuracy of maximum entropy distributions to ensembles of distributions that satisfy the same constraints, using the approach developed in Refs. [27,28].

II. MODEL ENSEMBLE

The collective state of a population of N variables is described by $\sigma = (\sigma_1, \ldots, \sigma_N)$. In general, σ_i may denote any degree of freedom, such as the identity of an amino acid in a protein, the orientation of a bird in a flock, etc. To fix this idea, in this paper σ_i will be a binary variable describing the spiking activity of neuron i: $\sigma_i = 1$ if neuron i spikes within a given time window, and 0 otherwise. The joint distribution of the collective activity σ, denoted by $P(\sigma)$, lives in a $2^N - 1$ dimensional space, represented schematically in Fig. 1. Because that space is huge for even moderately large populations, it is often impossible to sample the true distribution, $\hat{P}$, reliably from the data. Simplifying assumptions are needed.

To restrict the search of models, one can focus on distributions that agree with the data on the average value of a few observables. Calling these observables $\mathcal{O}_a(\sigma)$, $a = 1, \ldots, M$, the condition reads $P \cdot \mathcal{O}_a \equiv \sum_\sigma P(\sigma)\mathcal{O}_a(\sigma) = \overline{\mathcal{O}}_a$, where $\overline{\mathcal{O}}_a$ is the empirical mean. The observables must be chosen carefully depending on the problem at hand, and may include local or global order parameters, marginal probabilities, correlation functions, etc. Let us denote by $\mathcal{C}$ the subspace of models P that satisfy those constraints, as well as the conditions $P(\sigma) \geqslant 0$ and $\sum_\sigma P(\sigma) = 1$. $\mathcal{C}$ is convex because of the linear nature of the constraints.

A probability law on $\mathcal{C}$ may be defined, which weighs models $P \in \mathcal{C}$ according to their Gibbs entropy, $S(P) = -\sum_\sigma P(\sigma) \log P(\sigma)$, through the following measure [27]:

$$\mu_\Gamma(P) = \frac{e^{\Gamma S(P)}}{\mathcal{Z}}, \quad \mathcal{Z} = \int_{P \geqslant 0} \mathcal{D}P \, e^{\Gamma S(P)}, \tag{1}$$

*Corresponding author: tmora@lps.ens.fr

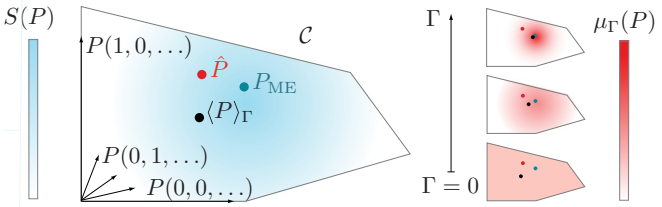

FIG. 1. Random models. The space $\mathcal{C}$ of models P is a simplex of $2^N - M - 1$ dimensions, defined by the intersection of the hyperplanes satisfying the constraints that the mean observables under the model, $\sum_\sigma P(\sigma)\mathcal{O}_a(\sigma)$, $a = 1,\ldots,M$, equal the empirical means, $\overline{\mathcal{O}}_a$, and by a normalization and positivity constraint. The true distribution to be approximated, $\hat{P}$ (red dot), is not accessible in general. An entropy-dependent measure μ_Γ [Eq. (1)] is defined on $\mathcal{C}$ (red map). At $\Gamma = 0$ (random ensemble), the measure is uniform over that space. As Γ is increased, the measure concentrates onto the maximum entropy distribution P_{ME} (blue dot), and so does its mean $P_\Gamma = \langle P\rangle_\Gamma$ (black dot).

with

$$\mathcal{D}P = \delta\left(\sum_\sigma P(\sigma) - 1\right)\prod_{a=1}^{M}\delta(P\cdot\mathcal{O}_a - \overline{\mathcal{O}}_a)\prod_\sigma dP(\sigma), \tag{2}$$

where $\delta(\cdot)$ is Dirac's δ function. The parameter Γ is conjugate to the entropy, and sets its average value: $\langle S(P)\rangle_\Gamma = \partial \ln \mathcal{Z}/\partial\Gamma$, where we use the brackets $\langle\cdot\rangle_\Gamma$ for averages over the measure μ_Γ. Γ plays the same role with respect to the ...s the inverse temperature with respect to the energy in ...ical mechanics. When $\Gamma = 0$, all distributions ...ints have the same probability. We ...ble. As $\Gamma \to \infty$, the measure ...single distribution, P_{ME}, ...alogy, the ground ...ion defines ...29]:

...ıl
...ased,
...aints.

...ED NEURAL ...ΓINA

...e random ensemble defined ...king activity of retinal ganglion ...There, the spiking activities of 40 ...e salamander retina were recorded by ...ys for about an hour, and segmented into ...ary spike words σ of 20 ms. The collective ...small networks (up to 10 neurons) was shown ...ell described by maximum entropy distributions ...ained by spike rates and pairwise correlations (and later to much larger populations [21]). This choice of constraints corresponds to the observables $\mathcal{O}_a = \sigma_i$ for all neuron i, and $\mathcal{O}_a = \sigma_i\sigma_j$ for all pairs i, j, for which the maximum entropy distribution (3) takes the form of a disordered Ising model, $P_{\mathrm{ME}}(\sigma) = (1/Z)\exp(\sum_i h_i\sigma_i + \sum_{ij} J_{ij}\sigma_i\sigma_j)$.

A. Small networks

It is instructive first to consider the unbiased measure μ_0 over very small networks, for which everything can be calculated analytically. The simplest case of two neurons constrained by just their firing rate is illustrated by Fig. 2(a). The maximum entropy distribution factorizes over the two neurons, which are thus independent [30]: $P(\sigma) = p_1(\sigma_1)p_2(\sigma_2)$. By contrast, random models drawn from μ_0 are biased towards a positive correlation $\langle\sigma_1\sigma_2\rangle - \langle\sigma_1\rangle\langle\sigma_2\rangle > 0$ when both firing rates $\langle\sigma_1\rangle, \langle\sigma_2\rangle$ are on the same side of 0.5 (in the retinal data $\langle\sigma_i\rangle \sim 0.02$). A similar bias in the triplet correlation is also found when considering three neurons constrained by uniform firing rates and pairwise correlations [Fig. 2(b)]. When pairwise correlations are weak, as is the case in the retina [3], random models predict on average a higher three-point connected correlation than maximum entropy, although the bias is reversed for large correlations.

B. Model comparisons

Thanks to its exponential form (3), the maximum entropy distribution can be inferred with relative ease for systems of size $N \leqslant 20$, yet requiring us to calculate sums of 2^N terms [3]. Sampling from μ_Γ or calculating P_Γ, on the other hand, is a much harder task, involving the exploration of $\mathcal{C}$ of dimension $2^N - N(N+1)/2 - 1$. To apply the random ensemble to populations of neurons, we sampled from μ_Γ with Monte Carlo using the Metropolis-Hastings algorithm,

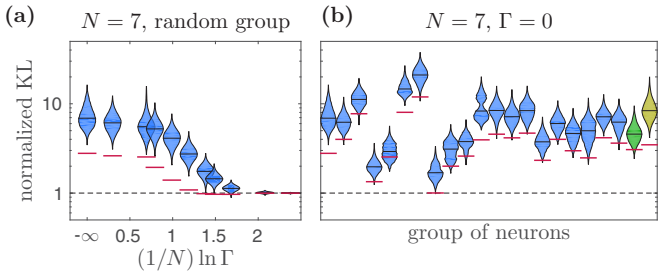

FIG. 3. Random versus maximum entropy models. (a) The normalized Kullback-Leibler divergence relative to maximum entropy, $D_{\mathrm{KL}}(\hat{P}\|P)/D_{\mathrm{KL}}(\hat{P}\|P_{\mathrm{ME}})$, is represented as a function of the entropy-conjugated variable Γ for (a) a random group of $N = 7$ neurons (out of 40). Values above unity (dashed line) mean that maximum entropy outperforms the random model. The violin plots show the distributions over random models drawn from μ_Γ, while the red lines show the value for the average model, $D_{\mathrm{KL}}(\hat{P}\|P_\Gamma)$. (b) Normalized KL divergence at $\Gamma = 0$ for 20 random subsets of seven neurons (blue), as well as the group of most correlated neurons (as measured by Pearson's correlation coefficient, green), and the set of neurons with the highest spike rate (yellow).

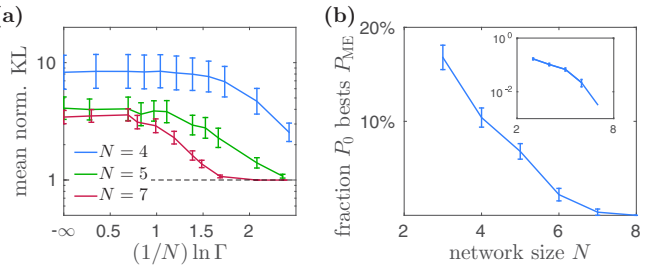

FIG. 4. Dependence on populations size. (a) The normalized divergence of the average model, $D_{\mathrm{KL}}(\hat{P}\|P_\Gamma)$, is averaged over 20 random subsets, and plotted as a function of $(1/N)\ln\Gamma$. Error bars show standard error on the mean. (b) Fraction of random groups (out of hundreds) of N neurons that are better described by the mean unbiased distribution P_0 than by the maximum entropy model P_{ME}.

for various subgroups of neurons of different sizes. At each step, starting from a distribution P in $\mathcal{C}$, one picks a random direction V in the Fourier basis of the hyperplane orthogonal to all observables $\mathcal{O}_a$ [28]. The new distribution is taken to be $P' = P + \alpha V$, where α is drawn uniformly in the interval $(\alpha_{\min},\alpha_{\max})$ defined by the lower and upper limits so that $P'(\sigma) \geqslant 0$ for all σ. P' is accepted with probability $\min(1, e^{\Gamma[S(P')-S(P)]})$. The process is repeated until equilibration is reached. High space dimension limits us to relatively small group sizes, $N \leqslant 8$. Fortunately for these sizes the true distribution $\hat{P}$ may be accurately estimated from the data, and directly compared to models.

The accuracy of a given model is assessed by the Kullback-Leibler (KL) divergence between the model distribution P and the true one $\hat{P}$, $D_{\mathrm{KL}}(\hat{P}\|P) = \sum_\sigma \hat{P}(\sigma)\ln[\hat{P}(\sigma)/P(\sigma)]$. Figure 3 shows, in the form of violin plots, the distribution of KL divergence (normalized relative to maximum entropy) when sampling P from μ_Γ, for groups of $N = 7$ cells. This distribution is plotted in Fig. 3(a) for a random group of seven cells. Maximum entropy is found to have a clear advantage: its accuracy is matched by only a negligible fraction of models drawn from μ_Γ, and it also does better than their mean P_Γ (red line). The advantage of maximum entropy over the unbiased ensemble generalizes to 20 random groups of seven cells [Fig. 3(b)], as well as the groups comprising the most correlated (green) and most active (yellow) cells. Interestingly, in all cases the mean distribution P_Γ is more accurate than the typical distribution P sampled from μ_Γ, a consequence of Jensen's inequality, which implies $D_{\mathrm{KL}}(\hat{P}\|\langle P\rangle_\Gamma) \leqslant \langle D_{\mathrm{KL}}(\hat{P}\|P)\rangle_\Gamma$. In general, $0 < \Gamma < \infty$ interpolates between the unbiased ensemble and the maximum entropy distribution. For these reasons, in the following the maximum entropy model P_{ME} will only be compared to the mean distribution of the unbiased ensemble, P_0. Note that estimating the mean distribution still requires us to sample $\mathcal{C}$ using Monte Carlo, and does not offer benefits in terms of computational speed.

C. Network size

We now investigate the dependence on the population size. Figure 4 shows the average normalized KL divergence of the mean model P_Γ for random cell groups of varying sizes, as a function of $(1/N)\ln\Gamma$ (the scaling of Γ is assumed to be exponential in N, as suggested by calculations with random observables [27]). The general trend noted before for $N = 7$ generalizes to all sizes: the larger the entropy bias Γ, the better the model [Fig. 4(a)]. However, this average behavior masks large heterogeneities across different choices of cell groups, especially for small groups, of which a sizable fraction is better described by the mean distribution P_0 than by P_{ME}. Evaluating this fraction from hundreds of random groups for each N, we find that maximum entropy is more likely to outperform the random ensemble in larger groups [Fig. 4(b)], and even does so in all of the 200 tested groups of size $N = 8$.

D. Maximum entropy and correlations

What sets apart groups of cells that are better described by P_0 than by P_{ME}? Since both share the same one- and two-point correlations by construction, we examine their predictions for three-point correlations in triplets of cells ($N = 3$). Figure 5(a) shows that random models typically fail because they overestimate small three-point correlations. By contrast, maximum entropy is more likely to be outperformed by random models when the triplet correlation is large, in which case maximum entropy overestimates it. Both these findings are in agreement with the results of Fig. 2(b). This observation can be generalized to larger groups of neurons ($N > 3$) by considering the total amount of correlations in the network, quantified by the loss of entropy due to correlations, or multi-information [30], $I = S(P_{\mathrm{ind}}) - S(\hat{P})$, where $P_{\mathrm{ind}} = \prod_i p_i(\sigma_i)$ is the model distribution of independent neurons. Groups that are better described by P_0 than by P_{ME} are found to have a higher multi-information on average [Fig. 5(c)].

IV. CONCLUSION

Since maximum entropy was proposed as a method for building statistical models from high-dimensional data, its accuracy, relevance, and epistemological validity have been questioned. In this study we have shown that the maximum

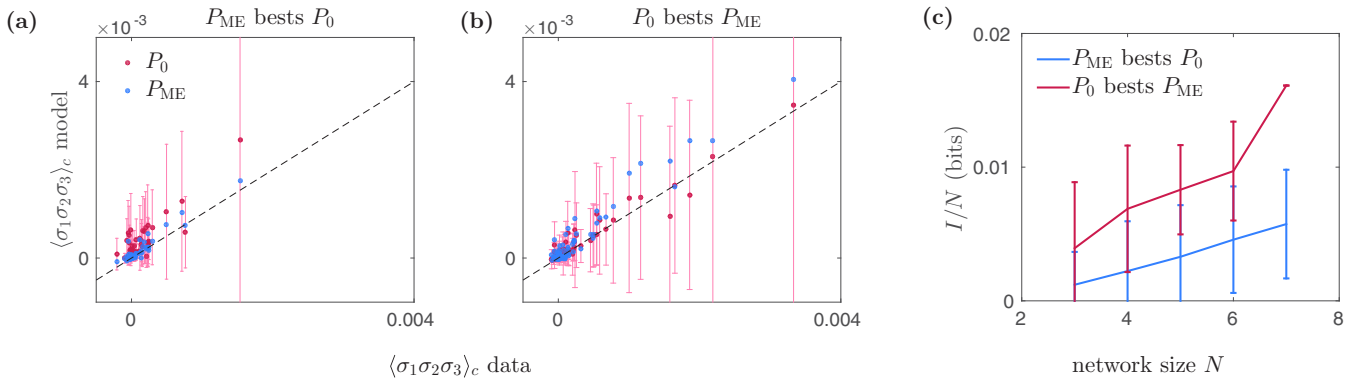

FIG. 5. Correlations and maximum likelihood performance. Three-point connected correlation $\langle\sigma_1\sigma_2\sigma_3\rangle_c$ for (a) 100 random triplets whose joint activity is best described by maximum entropy and (b) 100 random triplets whose joint activity is best described by the mean unbiased model, when constraining the values of the pairwise correlations. The error bar shows, for each triplet, the allowed range of values for the three-point correlation. (c) The multi-information, which measures the overall amount of correlation in the collective activity, is plotted as a function of system size, for groups of neurons that are best described by the maximum entropy model $P_{\rm ME}$ (blue) or by the mean unbiased model P_0 (red). Error bars show standard deviation across groups of cells (the red point at $N = 7$ has no error bar because only one group of that size was better described by P_0).

entropy model describes the spiking activity of populations in the retina better than the mean model satisfying the same constraints, which itself performs better than the vast majority of random models under these constraints. This better performance of maximum entropy gets more marked as the population size N grows, and is essentially always true for $N \geqslant 8$. The analysis of three-point and higher-order correlations suggests that the rare instances where the mean model outperforms maximum entropy involve relatively large correlations. When correlations are high, maximum entropy predicts high triplet correlations within the allowed range compared to the mean unbiased model [Fig. 2(b)], and may thus overestimate their true value, consistent with previous observations in large populations [21]. In that case, models that take a middle-of-the-road value of the correlations may be preferred to maximum entropy. We emphasize that although groups in which the mean model outperforms maximum entropy tend to have large correlations, the converse is not true: for instance all networks with $N \geqslant 8$ are better described by maximum entropy, regardless of their correlations.

It would be interesting to apply our approach to other data sets where maximum entropy has been shown to perform well, in particular on the correlated activity of cortical networks [5–7] where the nature of correlations may be different. Only such a comparative analysis could assess the general adequacy of maximum entropy beyond the particular case of the retina. However, by providing a first test on empirical data, our results complement previous work aimed at explaining or refuting the efficiency of maximum entropy solely based on theoretical arguments and simulated data sets.

In particular, the random ensemble of Eq. (1) was applied to synthetic models in which the observables were themselves picked at random as quenched disorder [27,28]. In Ref. [27], the form of the imposed constraints was drawn from a uniform distribution. Under that choice of observables, maximum entropy was found to be no more accurate than random. However, it is not clear how applicable these results are to real empirical distributions, and to pairwise constraints. For instance, if the constraints are not randomly selected, but chosen to reproduce lower-order statistics or to be smooth functions of the variables, the results could drastically change [28].

Other simulation studies have more specifically addressed the role of pairwise interactions. It was suggested that pairwise maximum entropy models should fail for large populations [31], but these conclusions were based on synthetic data simulated with higher-than-pairwise interactions, making pairwise maximum entropy models unfit almost by construction. On the other hand, strongly interacting systems with interactions of arbitrary order have been numerically shown to be well described by pairwise interactions, with an analogy to Hopfield networks [32]. The principle of maximum entropy has also been advocated by contrast to nonadditive (or Rényi) entropies, but on purely theoretical grounds [33] (although even more general classes of nonexponential distributions perform better, see Ref. [34]). Our results do not preclude that other objective functions than entropy may help better describe empirical data. They suggest, however, that for large networks it is better to pick the most random model than to pick a model at random.

ACKNOWLEDGMENTS

We thank Michael Berry for sharing the data from [3], and Olivier Marre for his comments on the manuscript. U.F. received funding from the European Union's Horizon 2020 research and innovation programme under Grant Agreements No. 604102 and No. 720270. T.O. was supported by KAKENHI No. 26870185.

U.F. and T.O. contributed equally to this work.

[1] E. T. Jaynes, Phys. Rev. **106**, 620 (1957).

[2] E. T. Jaynes, Phys. Rev. **108**, 171 (1957).

[3] E. Schneidman, M. J. Berry, R. Segev, and W. Bialek, Nature (London) **440**, 1007 (2006).

[4] J. Shlens, G. D. Field, J. L. Gauthier, M. I. Grivich, D. Petrusca, A. Sher, A. M. Litke, and E. J. Chichilnisky, J. Neurosci. **26**, 8254 (2006).
[5] A. Tang, D. Jackson, J. Hobbs, W. Chen, J. L. Smith, H. Patel, A. Prieto, D. Petrusca, M. I. Grivich, A. Sher *et al.*, J. Neurosci. **28**, 505 (2008).
[6] L. S. Hamilton, J. Sohl-Dickstein, A. G. Huth, V. M. Carels, K. Deisseroth, and S. Bao, Neuron **80**, 1066 (2013).
[7] G. Tavoni, U. Ferrari, F. P. Battaglia, and S. Cocco, Network Neurosci. (to be published).
[8] T. Watanabe, S. Hirose, H. Wada, Y. Imai, T. Machida, I. Shirouzu, S. Konishi, Y. Miyashita, and N. Masuda, Nature Commun. **4**, 1370 (2013).
[9] M. Weigt, R. A. White, H. Szurmant, J. A. Hoch, and T. Hwa, Proc. Natl. Acad. Sci. USA **106**, 67 (2009).
[10] T. Mora, A. M. Walczak, W. Bialek, and C. G. Callan, Proc. Natl. Acad. Sci. USA **107**, 5405 (2010).
[11] M. Figliuzzi, H. Jacquier, A. Schug, O. Tenaillon, and M. Weigt, Mol. Biol. Evol. **33**, 268 (2016).
[12] M. Santolini, T. Mora, and V. Hakim, PLoS One **9**, e99015 (2014).
[13] E. De Leonardis, B. Lutz, S. Ratz, S. Cocco, R. Monasson, A. Schug, and M. Weigt, Nucleic Acids Res. **43**, 10444 (2015).
[14] W. Bialek, A. Cavagna, I. Giardina, T. Mora, E. Silvestri, M. Viale, and A. M. Walczak, Proc. Natl. Acad. Sci. USA **109**, 4786 (2012).
[15] G. J. Stephens and W. Bialek, Phys. Rev. E **81**, 066119 (2010).
[16] E. D. Lee, C. P. Broedersz, and W. Bialek, J. Stat. Phys. **160**, 275 (2015).
[17] T. Mora and W. Bialek, J. Stat. Phys. **144**, 268 (2011).
[18] W. Bialek, A. Cavagna, I. Giardina, T. Mora, O. Pohl, E. Silvestri, M. Viale, and A. M. Walczak, Proc. Natl. Acad. Sci. USA **111**, 7212 (2014).
[19] G. Tkačik, T. Mora, O. Marre, D. Amodei, S. E. Palmer, M. J. Berry, and W. Bialek, Proc. Natl. Acad. Sci. USA **112**, 11508 (2015).
[20] T. Watanabe, N. Masuda, F. Megumi, R. Kanai, and G. Rees, Nature Commun. **5**, 4765 (2014).
[21] G. Tkačik, O. Marre, D. Amodei, E. Schneidman, W. Bialek, and M. J. Berry, PLoS Comput. Biol. **10**, e1003408 (2014).
[22] F. Morcos, A. Pagnani, B. Lunt, A. Bertolino, D. S. Marks, C. Sander, R. Zecchina, J. N. Onuchic, T. Hwa, and M. Weigt, Proc. Natl. Acad. Sci. USA **108**, E1293 (2011).
[23] A. L. Ferguson, J. K. Mann, S. Omarjee, T. Ndung'u, B. D. Walker, and A. K. Chakraborty, Immunity **38**, 606 (2013).
[24] J. E. Shore and R. W. Johnson, IEEE Trans. Inf. Theory **26**, 26 (1980).
[25] E. Aurell, PLoS Comput. Biol. **12**, e1004777 (2016).
[26] E. van Nimwegen, PLoS Comput. Biol. **12**, e1004726 (2016).
[27] T. Obuchi, S. Cocco, and R. Monasson, J. Stat. Phys. **161**, 598 (2015).
[28] T. Obuchi and R. Monasson, J. Phys.: Conf. Ser. **638**, 012018 (2015).
[29] S. Presse, K. Ghosh, J. Lee, and K. A. Dill, Rev. Mod. Phys. **85**, 1115 (2013).
[30] E. Schneidman, S. Still, M. J. Berry, and W. Bialek, Phys. Rev. Lett. **91**, 238701 (2003).
[31] Y. Roudi, S. Nirenberg, and P. E. Latham, PLoS Comput. Biol. **5**, e1000380 (2009).
[32] L. Merchan and I. Nemenman, J. Stat. Phys. **162**, 1294 (2016).
[33] S. Pressé, K. Ghosh, J. Lee, and K. A. Dill, Phys. Rev. Lett. **111**, 180604 (2013).
[34] J. Humplik and G. Tkačik, arXiv:1605.07371 (unpublished).