A Study of the Boltzmann Sequence-Structure Channel

Abram Magner and Daisuke Kihara and Wojciech Szpankowski, *Fellow, IEEE,*

***Abstract*—We rigorously study a channel that maps binary sequences to self-avoiding walks in the two-dimensional grid, inspired by a model of protein folding from statistical physics and studied empirically by biophysicists. This channel, which we also call the Boltzmann sequence-structure channel, is characterized by a Boltzmann/Gibbs distribution with a free parameter corresponding to temperature. In our previous work, we verified experimentally that the channel capacity appears to have a phase transition for small temperature and decays to zero for high temperature. In this paper, we make some progress towards explaining these phenomena. We first estimate the conditional entropy between the input sequence and the output fold, giving an upper bound which exhibits a phase transition with respect to temperature. Next, we formulate a class of parameter settings under which the dependence between walk energies is governed by their number of shared contacts. In this setting, we derive a lower bound on the conditional entropy. This lower bound allows us to conclude that the mutual information tends to zero for high temperature, giving some support to the experimental fact regarding capacity which tends to zero in this regime. Finally, we construct an example setting of the parameters of the model for which the free energy is exactly calculable.**

I. Introduction

Information theory traditionally deals with the problem of transmitting sequences over a communication channel and finding the maximum number of messages that the receiver can recover with arbitrarily small probability of error. However, databases of various sorts have come into existence in recent years that require to transmit structural data (e.g., graphs and sets). Contemporaneously, there has been significant effort focused on understanding the equilibrated states and dynamics of biomolecules [1], in particular, to determine folded states and fold changes. We bridge these seemingly disparate ideas using novel information theoretic modeling. In [2], we attempted an information-theoretic explanation of a few observations previously made by biophysicists: while the number of amino acid sequences observed in nature is large, the corresponding number of dissimilar tertiary structures to which the sequences have been observed to fold is relatively small. Additionally, the frequency distribution of protein families observed in nature exhibits power law characteristics. We provided experimental evidence that explains these observations by modeling the protein folding process as a channel. We gave evidence in support of the hypothesis that these complex phenomena might have interesting information theoretic underpinnings.

This channel maps binary (hydrophobic, denoted by H, and polar, denoted by P) sequences into two-dimensional self-avoiding walks (also called folds) in a square lattice (see Figure 1). * A sequence of length N induces a labeling of each fold of the same length, and counting the number of different types of contacting nodes induces an energy function on the set of folds. This energy function induces a conditional probability distribution on the set of folds, where lower energy folds receive higher probability.

In particular, the channel is defined by the Boltzmann/ Gibbs distribution with a free parameter corresponding to inverse temperature. We therefore call it the *Boltzmann sequence-structure channel*. For such a channel, the key parameter is the conditional entropy between the input sequence and the output fold. In this paper, we provide a mathematically rigorous foundation to estimate this entropy and show that it may exhibit a range of interesting behaviors with respect to temperature, depending on the settings of the parameters of the model.

Fig. 1: A sequence passing through the channel and being paired with a fold given by a self-avoiding walk.

We now describe in more detail the construction of the channel. For each sequence s, the folds f are assigned energies $\mathcal{E}(f, s)$ depending on the number of different types of *contacts* between residues, that is, between neighboring, but not sequence-adjacent, nodes of the self-avoiding walk. These

A. Magner (email: anmagner@purdue.edu), D. Kihara (email: dkihara@purdue.edu), and W. Szpankowski (email: spa@cs.purdue.edu) are with the Department of Computer Science at Purdue University, West Lafayette, IN, USA. D. Kihara is also with the Department of Biological Sciences. W. Szpankowski is also with the Faculty of Electronics, Telecommunications and Informatics, Gdańsk University of Technology, Poland. This work was supported by NSF Center for Science of Information (CSoI) Grant CCF-0939370, and in addition by NSF Grants CCF-1524312, and NIH Grant 1U01CA198941-01, and the NCN grant, grant UMO-2013/09/B/ST6/02258.

*We discuss below and in the literature review the history/justification (given by biophysicists) of the classification of amino acids into hydrophobic and polar, as well as the modeling of protein structures and more general polymers as self-avoiding walks in a lattice.

contact energies are given by a *scoring* matrix Q whose rows and columns are indexed by H and P. Since hydrophobic interactions are a dominant force for protein folding, it is reasonable to classify amino acids into hydrophobic (H) and polar (P). Thus, in a realistic lattice model, contacts between H and H are more favored (lower energy) than H and P interactions [3]. The channel is then defined by the Boltzmann distribution induced by the energies.

More precisely, for each even (for technical reasons explained below) perfect square integer N, we have an input set $\mathcal{S}_N$ consisting of 2^N sequences of length N over the alphabet $\{H, P\}$. The output set $\mathcal{F}_N$ consists of all directed self-avoiding walks of length N on a $\sqrt{N}\times\sqrt{N}$ integer lattice which start at $(0,0)$ and end at $(\sqrt{N}-1, \sqrt{N}-1)$. Note that all but $O(\sqrt{N})$ points in the lattice have four neighbors (but only two contact points) since every walk fills the lattice completely. We endow each sequence/fold pair with an energy as follows: fix a symmetric 2×2 matrix $Q = \{Q_{ij}\}_{i,j\in\{1,2\}}$ over $\mathbb{R}$ (the scoring matrix). For $f \in \mathcal{F}_N$ and $s \in \mathcal{S}_N$

$$\mathcal{E}(f,s) = 2(Q_{11}c_{HH} + Q_{22}c_{PP} + Q_{12}c_{HP}), \qquad (1)$$

where c_{xy} denotes the number of contacts $\{a,b\}$ such that $s_a = x$ and $s_b = y$ or vice-versa (throughout, for any sequence s and $j \in [N] = \{1,\dots,N\}$, we denote by s_j the jth symbol of s). Here, the multiplication by 2 is for mathematical convenience and is insignificant to the analysis. Then we define the channel by the conditional probability $p_N(f|s)$ that follows the Boltzmann distribution.

More formally, let $\beta \geq 0$ be a real number (corresponding to an inverse temperature). Then we define

$$p_N(f|s) = p(f|s) = \frac{e^{-\beta\mathcal{E}(f,s)}}{Z(s,\beta)}, \quad Z(s,\beta) = \sum_{f\in\mathcal{F}_N} e^{-\beta\mathcal{E}(f,s)},$$

where the function Z is known as the *partition function*, which plays a central role in statistical mechanics models as a kind of generating function of configuration energies. Two quantities will play an especially important part in our analysis and results: the free energy $\gamma_N(\beta)$ is given by

$$\gamma_N(\beta) = \frac{\mathbb{E}\log Z(S,\beta)}{\log|\mathcal{F}_N|} \qquad \gamma(\beta) = \limsup_{N\to\infty}\gamma_N(\beta).$$

We also denote by μ the exponential growth rate of the number of self-avoiding walks:

$$\mu_N = \frac{\log|\mathcal{F}_N|}{N}, \qquad \mu = \lim_{N\to\infty}\frac{\log|\mathcal{F}_N|}{N}.$$

Both are challenging to compute.

This channel is interesting from the information-theoretic point of view, irrespective of applications, primarily because it exhibits several unusual mathematical properties: first, it maps sequences to structures (i..e, self-avoiding walks) in a nontrivial way; second, it is a channel with full memory; and, finally, several information theoretic quantities associated with it (e.g., its capacity and conditional entropy for certain natural input distributions) likely exhibit phase transitions with respect to temperature for certain settings of the scoring matrix. Probabilistically, its analysis presents an interesting challenge because the nontrivial dependence structure between fold energies makes bounding the variance of the number of folds with a given maximum energy difficult. This in turn, complicates the calculation of the free energy, which plays a significant role in our calculations (and, for many models, is notoriously difficult to compute [4]). Since the exponential growth rate of the number of folds in the output alphabet appears in several quantities of interest, we also encounter combinatorial problems which are currently under active investigation.

We now review some of the relevant literature.

Regarding self-avoiding walks (*SAWs*), [5] is a good general reference, including a discussion of the history of the use of SAWs as models for polymers. SAWs continue to be used as simple models for protein structures in molecular biology (see, e.g., [6], [7]). One of the fundamental problems in the theory of SAWs is the (asymptotic) enumeration of classes $\mathcal{F}_N$ of SAWs of length $N \to \infty$ with various constraints. In particular, the problem of proving the existence/determining the value of the limit

$$\lim_{N\to\infty}|\mathcal{F}_N|^{1/N}$$

(called the *connective constant* of $\mathcal{F}_N$) is commonly studied and is quite challenging. There are a few general techniques for approaching such problems, sub/superadditivity arguments being the main ones. For, say, subadditivity, the goal is to show that, for all $1 \leq m \leq N-1$,

$$|\mathcal{F}_N| \leq |\mathcal{F}_m||\mathcal{F}_{N-m}|, \qquad (2)$$

which implies that the sequence $(\log|\mathcal{F}_N|)_{N=1}^{\infty}$ is subadditive. By, e.g., Fekete's lemma (or one of its generalizations) [8], this is sufficient to conclude the existence of the limit

$$\lim_{N\to\infty}\frac{\log|\mathcal{F}_N|}{N}.$$

Usually, the condition (2) can be verified by some sort of splitting (or concatenation, in the case of superadditivity) in order to establish an injection from $|\mathcal{F}_N|$ to $\mathcal{F}_m \times \mathcal{F}_{N-m}$. For example, if we take $\mathcal{F}_N$ to be the set of all SAWs, we can split a walk $w \in \mathcal{F}_N$ into a unique initial part of length m and a final part of length $N-m$, which establishes (2). In general, determining the value of the connective constant requires significant ingenuity (see, e.g., [9], which establishes the value for SAWs on the two-dimensional hexagonal lattice).

Even proving/disproving the existence of a connective constant becomes significantly harder when we consider collections of SAWs satisfying some geometric constraints (unless they are very carefully chosen). For instance, consider the set of Hamiltonian SAWs filling a square of size N (with N a perfect square). Neither splitting nor concatenation works

here, since neither operation yields SAWs within the same class in general. By adding the constraint that each SAW must begin at a fixed corner of the square and end at the opposite and restricting to an appropriate subsequence (i.e., even and perfect square N), [10] showed the existence of the connective constant as a limit of that subsequence (though the result is incorrectly stated; see [11] for a discussion and estimates of the limit).

We now review what is known about some relevant models from statistical physics. For general references, see [4], [12]. For a set Γ_N of configurations, each configuration $\xi \in \Gamma_N$ is endowed with its own (possibly random) *energy* $\mathcal{E}(\xi)$. The set Γ_N is then endowed with a probability distribution governed by this energy (chosen so as to have maximum entropy under the constraint that the system has a given energy density), known as the *Boltzmann/Gibbs* measure:

$$p(\xi) = \frac{e^{-\beta \mathcal{E}(\xi)}}{Z(\beta)},$$

where $\beta \in [0, \infty)$ is a free parameter which intuitively behaves like an inverse temperature, and Z above is the *partition function*, given by

$$Z(\beta) = \sum_{\xi \in \Gamma_N} e^{-\beta \mathcal{E}(\xi)}.$$

The main problem is to establish the existence/estimate the asymptotic value of the *free energy*:

$$\lim_{N \to \infty} \frac{\mathbb{E}[\log Z(\beta)]}{\log |\Gamma_N|}. \tag{3}$$

This quantity is studied because other important parameters, such as the *entropy density* and *energy density* can be written in terms of it (see [12] for details). One of the simplest interesting models is the *random energy model* (REM), in which the configuration space has size 2^N, and the configurations are i.i.d. (exactly) Gaussian random variables: $\mathcal{E}(\xi) \sim \mathcal{N}(0, N/2)$. The free energy for this model is exactly solvable (which is unusual for these sorts of models):

$$\lim_{N \to \infty} \frac{\mathbb{E}[\log Z(\beta)]}{N} = \begin{cases} \beta^2/4 + \log 2 & \beta \leq 2\sqrt{\log 2} \\ \beta \sqrt{\log 2} & \beta \geq 2\sqrt{\log 2}. \end{cases}$$

Note that the free energy exhibits a phase transition with respect to temperature, since, for small β, it grows quadratically, while it grows linearly when $\beta \geq 2\sqrt{\log 2}$. This sort of phenomenon is quite common (though not universal) in statistical physics, and we will encounter it in our analysis in this paper.

The situation becomes significantly more complicated when correlations between configurations are introduced. For instance, in the Sherrington-Kirkpatrick (SK) model, configurations are strings of length N from the alphabet $\{-1, 1\}$, and the energy of a configuration ξ is given by

$$\mathcal{E}(\xi) = -\frac{\beta}{\sqrt{N}} \sum_{i<j} g_{ij} \xi_i \xi_j,$$

with i.i.d. random variables $g_{ij} \sim \mathcal{N}(0, 1)$. The correlation between two configurations $\xi^{(1)}$ and $\xi^{(2)}$ then increases with the number of indices i for which $\xi_i^{(1)} = \xi_i^{(2)}$. This model was introduced in [13], in which the authors also gave an incorrect expression for the free energy. Parisi, in [14], conjectured the correct formula (which now bears his name), but over 20 years passed before it was rigorously verified by Talagrand in [15].

We now move on to discuss our contributions. First, though the self-avoiding walk model and associated energy function for proteins has been considered empirically before [6], [7], we appear to be the first to define the channel that we consider here and study its information theoretic quantities. Of particular interest is the *capacity* of the channel:

$$C = \max_{p(S)} [H(F) - H(F|S)],$$

where the maximum is taken over all probability distributions on the set of sequences; see [16]. In our previous work [2], we studied this quantity numerically. Specifically, using a specific scoring matrix taken from the biology literature, we computed the conditional probabilities constituting the channel for $N = 36$ (due to computational limitations, we could not do the same for much larger N). We then computed the capacity for various temperatures using the Blahut-Arimoto algorithm ([16]), resulting in Figure 2. We note two phenomena illustrated by the plot: first, there appears to be a phase transition with respect to temperature in the capacity. Second, the capacity tends to 0 as temperature tends to infinity (for fixed N, this is simple to prove, but significantly more interesting when $N \to \infty$).

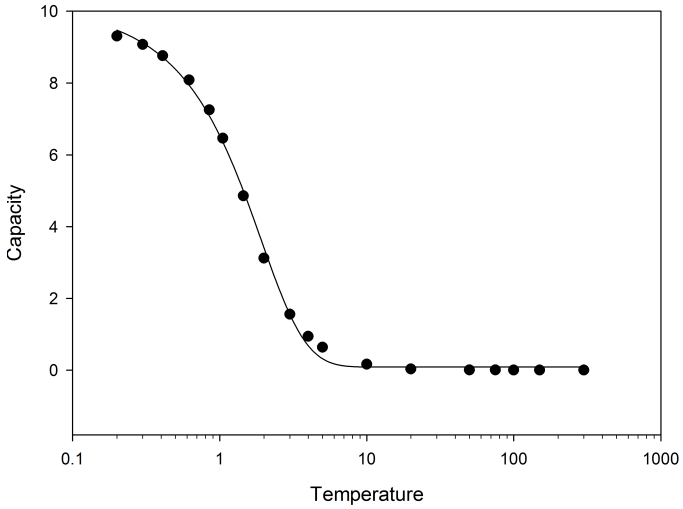

Fig. 2: Empirical evidence of a phase transition in channel capacity. Here, the capacity at various temperatures for the channel associated with 6×6 lattices is depicted. See [2] for the full figure.

As a long-term goal, we would like to rigorously establish the asymptotic behavior of the capacity of this channel for all

temperatures and suitable scoring matrices. Our focus in this work is more modest: we mainly study here the behavior of the conditional entropy for a memoryless source in the high-temperature regime (i.e., $\beta \xrightarrow{N\to\infty} 0$).

First, we give upper bounds on the free energy (hence the conditional entropy) whose behavior depends on the difference between the expected energies of a Boltzmann-distributed fold and one chosen uniformly at random. We then show how a series representation, involving the higher moments of the partition function, may be derived for the free energy via Taylor's expansion. Next, we present a class of scoring matrices for which the covariance between any two fold energies depends on the number of shared contacts between the two folds. For such matrices, we derive a formula for the variance of the partition function in terms of the number of contacts shared between two random folds, which implies a lower bound on the free energy. As an application of the lower bound, we give a sufficient condition on the temperature under which the mutual information between the channel input and output tends to 0. Finally, we point out that the model may exhibit a diverse range of behaviors depending on parameter settings by exhibiting a class of scoring matrices for which the free energy is exactly analyzable and has capacity $o(N)$ for any β.

The model presents several mathematical challenges: due to geometric constraints (e.g., Hamiltonicity), the configurations (folds) cannot easily be decomposed into subconfigurations. Thus, techniques which are useful for other models (e.g., [17]) do not appear to be easily adapted to our case. Probabilistically, the correlation structure between fold energies does not appear to be captured by other existing models (e.g., the REM, the *generalized REM* (GREM) [18], or the SK model). Moreover, while many models are defined so that configuration energies are normally distributed, the fold energies are only *asymptotically* normally distributed. Finally, our analysis leads to some classic open questions about enumerating self-avoiding walks, including proving the existence of the connective constant for geometrically constrained walk sets and determining distributional information about the number of shared contacts between two randomly chosen folds.

II. Main results

We now fix some useful notation, precisely describe the model, and state our main results.

A. Description of the model

Throughout, we use F to denote a *random* fold generated by choosing a random sequence according to some distribution and passing it through the channel. We generally use f to denote an arbitrary fixed fold. For any fold $f \in \mathcal{F}_N$, we denote the two-dimensional position of the jth node in f by $\pi_f(j)$. For any $j, k \in [N]$, we say that j and k are *sequence-adjacent* if $|j-k| = 1$ (here, $[N] = \{1, 2, \ldots, N\}$). We say that they are *lattice-adjacent* and that they form a *contact* if they are not sequence-adjacent and $\|\pi_f(j) - \pi_f(k)\|_1 = 1$ (here, $\|\cdot\|_1$ denotes the ℓ_1 norm on $\mathbb{Z}^2$). This allows us to define the energy $\mathcal{E}(f,s)$ as in (1). We also define $\mathcal{E}_{\beta,S}(F)$ to be the energy of the fold generated by the channel at inverse temperature β with the sequence S on its input.

We can also express the $\mathcal{E}(f,s)$ as a sum of *local energies*: for each $i \in [N]$, define $X_i = X_i(f,s)$ to be

$$X_i = Q_{11}c_{HH}(i) + Q_{22}c_{PP}(i) + Q_{12}c_{HP}(i),$$

where $c_{xy}(i)$, discussed above, denotes the number of contacts $\{i,j\}$ whose sequence elements are x and y or vice-versa (we note that the multiplication by 2 in (1) is because, by summing over all X_i, we count each contact twice). Then we have

$$\mathcal{E}(f,s) = \sum_{i=1}^{N} X_i(f,s).$$

Clearly,

$$\mathbb{E}[\mathcal{E}(f,S)] = \sum_i \mathbb{E}[X_i(f,S)] = N\alpha + O(\sqrt{N})$$

for some easily computable α depending on Q (with $\alpha \neq 0$ under mild conditions on Q and the sequence distribution), where boundary conditions contribute the $O(\sqrt{N})$. In fact, we can give an explicit formula for α:

$$\alpha/2 = p^2 Q_{HH} + 2pqQ_{HP} + q^2 Q_{PP}.$$

In contrast, $\mathbb{E}[\mathcal{E}_{\beta,S}(F)]$, the expected energy of a Boltzmann fold, is more difficult to compute. We discuss some of its properties below.

We restrict our attention to a particular class of distributions on $\mathcal{S}_N$ that is natural to consider: the symbols are i.i.d. random variables, taking the value H with probability $p \in (0,1)$ and P with probability $q = 1-p$. That is, we take a binary memoryless source with parameter p, which we denote by $\mathcal{B}_N(p)$. Many of our results can be extended to more general mixing sources.

As mentioned earlier, we restrict our attention to the class of Hamiltonian SAWs on a square, starting at the origin and ending at the opposite corner, and we restrict to N for which $\mathcal{F}_N$ is nonzero.

B. Statement of main results

We start with an expression for the conditional entropy. We have

$$\begin{aligned} H(F|S) &= -\sum_{s\in\mathcal{S}_N} p(s) \sum_{f\in\mathcal{F}_N} p(f|s)\log p(f|s) \\ &= \mathbb{E}[\log Z(S,\beta)] + \beta \sum_{s,f} p(f,s)\mathcal{E}(f,s) \\ &= \mathbb{E}[\log Z(S,\beta)] + \beta\mathbb{E}[\mathcal{E}_{\beta,S}(F)] \end{aligned} \tag{4}$$

where $\mathcal{F}_N$ denotes a set of self-avoiding walks of length N and we explicitly write

$$\mathbb{E}[\mathcal{E}_{\beta,S}(F)] = \sum_{s,f} p(f,s)\mathcal{E}(f,s).$$

The first and third equalities are elementary, and the second is by substitution of the definition of the channel into the right-hand side. Dividing by N on both sides, we have

$$\frac{H(F|S)}{N} = \frac{\log|\mathcal{F}_N|}{N} \cdot \frac{\mathbb{E}[\log Z(S,\beta)]}{\log|\mathcal{F}_N|} + \beta\frac{\mathbb{E}[\mathcal{E}_{\beta,S}(F)]}{N}.$$

It is easy to see that $\mathbb{E}[\log Z(S_N,\beta)] = O(N)$, so that the free energy $\gamma(\beta) < \infty$. Moreover, defining

$$\alpha_*(\beta,N) = \alpha_*(\beta) = \frac{\mathbb{E}[\mathcal{E}_{\beta,S}(F)]}{N},$$

it can be shown that $\alpha_*(\beta) < \infty$ for all β.

We note an important property of $\mathbb{E}[\mathcal{E}_{\beta,S}(F)]$: for an arbitrary fold f (equivalently, a uniformly distributed fold f, since both have the same expected energy when labeled by a sequence from a memoryless source)

$$\mathbb{E}[\mathcal{E}_{\beta,S}(F)] \le \mathbb{E}[\mathcal{E}(f,S)]. \tag{5}$$

This follows from an easy inductive proof, using the fact that the Boltzmann energy distribution is monotone decreasing (i.e., the Boltzmann distribution gives higher probability to lower energy folds).

We have the following upper bound on the free energy, and, hence, the conditional entropy.

Theorem 1 (Upper bound on the conditional entropy for memoryless sources)**.** *For any distribution over $\mathcal{S}_N$, $\beta > 0$, and scoring matrix Q,*

$$\frac{H(F|S)}{N} = \mu \cdot \gamma_N(\beta) + \beta\alpha_*(\beta) + o(1). \tag{6}$$

Furthermore, when $S \sim \mathcal{B}_N(p)$, if the scoring matrix Q is such that, uniformly over all $f \in \mathcal{F}_N$,

$$\mathbf{Var}\ [\mathcal{E}(f,S)] \sim N\sigma^2,$$

with $\sigma > 0$ constant with respect to N, then we have the following upper bound: for all $\beta > 0$,

$$\frac{H(F|S)}{N} \le \mu_N - \beta(\alpha - \alpha_*(\beta)) + \frac{1}{2}\sigma^2\beta^2 - O(\beta N^{-1/2}), \tag{7}$$

with $\mu = \lim_{N\to\infty}\mu_N$, and for bounded $\beta \ge \beta_* = \frac{\sqrt{2\mu}}{\sigma}$,

$$H(F|S) \le \beta N(\sqrt{2\sigma^2\mu_N} - (\alpha - \alpha_*(\beta)) + O(N^{-1/2})), \tag{8}$$

with the threshold value $\beta_* = \frac{\sqrt{2\mu}}{\sigma}$.

The condition on the scoring matrix is quite general. It is equivalent to requiring that Q_{HH}, Q_{HP}, and Q_{PP} are not all equal (in this case, a typical contact energy has positive variance).

Remark There is an information-theoretic upper bound on $H(F|S)$:

$$H(F|S) \le H(F) \le \log|\mathcal{F}_N| = N\mu_N$$

Provided that $\beta = o(1)$ and $\alpha - \alpha_*(\beta) = \Theta(1)$, the first upper bound given above beats this one. Similarly, if $\alpha - \alpha_*(\beta)$ is sufficiently large for any fixed β, the second upper bound is nontrivial. Moreover, the proof of the second upper bound implies that a refinement of the first upper bound for all β yields a corresponding refinement in the second.

Our next theorem gives, for each $p \in (0,1)$, a natural class of scoring matrices that endows the set of fold energies with a correlation structure similar to that arising in several models associated with combinatorial optimization problems (see [19]). In particular, the covariance between the energies of two folds f and g varies linearly with a measure of *overlap* between them: namely, the number of shared contacts between f and g (denoted by $k_{f,g}$). For such matrices, we establish a lower bound which holds for sufficiently small β, depending on the behavior of the MGF of the random variable K (the number of shared contacts between two folds chosen uniformly at random with replacement).

Theorem 2 (Free energy lower bound for high temperature)**.** *Let $S \sim \mathcal{B}_N(p)$ for fixed $p \in (0,1)$. Let K denote the number of shared contacts between two folds $f, g \in \mathcal{F}_N$ chosen uniformly at random with replacement. There exists a scoring matrix for which, provided that*

$$\mathbb{E}_K[e^{3\beta_N^2\sigma^2 K}] = 1 + o(1), \tag{9}$$

and $\beta = \beta_N = o(1)$, we have

$$\frac{H(F|S)}{N} \ge \mu_N - \beta(\alpha - \alpha_*(\beta)) + \frac{1}{2}\beta^2\sigma^2 - o(1), \tag{10}$$

where the $o(1)$ is expressible in terms of $\mathbb{E}_K[e^{3\beta_N^2\sigma^2 K}]$.

We remark that while essentially nothing is known about K in the condition (9), we do know that $K \le N + O(\sqrt{N})$, since that is the total number of contacts in a fold. Thus, a sufficient condition for (9) to hold is that $\beta = o(N^{-1/2})$. However, since we suspect that $K = O(1)$ with high probability, it seems likely that this can be relaxed. Note that the lower bound (10) matches the upper bound (7) up to the β term if $\alpha - \alpha_*(\beta) = \Theta(1)$ and the $o(1)$ term is $o(\beta)$.

Also, note that one cannot expect such a lower bound for general scoring matrices. This is because, for "most" matrices, the covariance of the energies of two contacts (i.e., unordered pairs of distinct sequence indices) which share exactly one node is positive, which implies that the covariance between two node energies is positive. This, in turn, implies that the covariance between *any* two fold energies is linear in N; that is, the dependence between fold energies is quite strong, in contrast with the situation in the REM. The scoring matrices considered in Theorem 2 are chosen so that the covariance

between energies of nonidentical contacts is 0, so that the covariance between folds is only linear in the number of shared contacts.

Remark For $\beta \xrightarrow{N\to\infty} 0$ and the class of scoring matrices considered in Theorem 2, we may be able to refine our estimate of the coefficient of β^2 in the expansion of the free energy by Taylor expanding the function $\log Z$ around $Z = \mathbb{E}[Z]$ and then taking expectations [19]:

$$\mathbb{E}[\log Z] = \log \mathbb{E}[Z] - \frac{\mathbf{Var}\ [Z]}{2(\mathbb{E}[Z])^2} \tag{11}$$

$$+ \sum_{m=3}^{\infty} \frac{(-1)^{m+1}}{m} \cdot \frac{\mathbb{E}[(Z - \mathbb{E}[Z])^m]}{(\mathbb{E}[Z])^m}. \tag{12}$$

This boils the problem down to the estimation of the centered moments of the partition function. For example, according to Lemma 4, used in the proof of Theorem 2 above, and Lemma 3, the first two terms of the expansion (11) are

$$\log |\mathcal{F}_N| - \beta\alpha N + \frac{1}{2}\beta^2\sigma^2 N$$
$$- (1 + O(N^{-1/2}))(\mathbb{E}_K[e^{3\beta^2\sigma^2 K}] - 1)/2 + O(N^{-1/2}).$$

Provided that $\beta = o(N^{-1/2})$, the contribution of the variance term becomes asymptotically equivalent to

$$-3\beta^2\sigma^2\mathbb{E}[K]/2.$$

In particular, note that both the expected value and variance terms of (11) contribute to the coefficient of β^2. More generally, the mth moment may be written in terms of the MGFs of the random variables $K_{m,j}$, for $j = 1, \ldots, m$, defined to be the number of contacts shared among exactly j folds among m folds chosen uniformly at random with replacement. The random variable K is a special case: $K = K_{2,2}$.

Provided that $K_{m,j}$ are sufficiently well-behaved, the series (11) above converges, and this gives a series representation for the coefficient of β^2, which may be bounded.

Depending on the asymptotics of the difference $\alpha - \alpha_*(\beta)$, Theorem 2 yields an interesting result about the mutual information $I(F;S) = H(F) - H(F|S)$ as the temperature tends to ∞. When α and $\alpha_*(\beta)$ are asymptotically equivalent and β is sufficiently small, the lower bound of Theorem 2 implies that $H(F|S) = \log|\mathcal{F}_N| - o(1)$. Thus, $I(F;S) = o(1)$.

Corollary 1. *With p and the scoring matrix Q as in Theorem 2, if β_N is such that $\alpha = \alpha_*(\beta_N) + \psi(N)$, where $\psi(N) = o(1)$ and $\beta_N\psi(N)N = o(1)$, and $\beta_N = o(N^{-2/3})$, then*

$$I(F;S) = o(1). \tag{13}$$

Note that one naturally expects that the mutual information tends to 0 when the temperature tends to infinity quickly enough (because then the Boltzmann distribution converges to uniformity), but this only becomes trivial when $\beta_N = O(1/N)$. The corollary, being a statement about the decay of the mutual information, is a small step in the direction of our stated goal of characterizing the capacity of the channel, in particular determining when it tends to 0.

We next give an example scoring matrix which exhibits a rather different behavior from the ones in Theorem 2.

Theorem 3 (An exactly analyzable scoring matrix with no phase transition)**.** *Let Q be the scoring matrix which maps $HH \mapsto -1/2$, $HP/PH \mapsto -1/4$, $PP \mapsto 0$. Then, for arbitrary sequence distributions, the free energy is given by*

$$\gamma(\beta) = 1 + \beta \limsup_{N\to\infty} \frac{\mathbb{E}[D_S(H)]}{\log|\mathcal{F}_N|},$$

where $D_S(H)$ is the number of i for which $S_i = H$. In the case of $S \sim \mathcal{B}_N(p)$, this becomes

$$\gamma(\beta) = 1 - \beta\alpha/\mu.$$

This theorem gives an example of a natural scoring matrix for which there is no first-order phase transition in the free energy. Moreover, it gives an upper bound on any lower bound for all (or almost all) scoring matrices that we can hope to prove.

III. Proofs

A. Proof of Theorem 1

The general expression for the asymptotic conditional entropy was already derived, so we give here the proof of the upper bounds. To do this, we prove analogous bounds for the free energy. In deriving the first bound, we will use Jensen's inequality to bring the expectation within the logarithm in the definition of the free energy. This will result in an expression involving the MGFs $\phi_N(\cdot)$ of appropriately normalized fold energies, which we will show to be asymptotically equivalent to the MGFs $\phi(\cdot)$ of Gaussian random variables with the same mean and variance. This is nontrivial, since a central limit theorem only a priori implies that $\phi_N(t) \xrightarrow{N\to\infty} \phi(t)$ for fixed $t \in \mathbb{R}$, whereas we need asymptotics for $\phi_N(t\sqrt{N})$.

We start by showing that fold energies are asymptotically normally distributed.

Lemma 1 (Fold energy CLT)**.** *Let $S_N \sim \mathcal{B}_N(p)$ for fixed $p \in (0,1)$. Let, for any $f \in \mathcal{F}_N$,*

$$\hat{\mathcal{E}}_N = \frac{\mathcal{E}(f, S_N) - \mathbb{E}[\mathcal{E}(f, S_N)]}{\sqrt{N}},$$

and denote by $F_N(\cdot)$ the distribution function of $\hat{\mathcal{E}}_N$. There exists a polynomial $V(p)$ whose coefficients are polynomials in the entries of the scoring matrix Q, such that, provided V is not identically zero, for all but finitely many choices of p, $\sigma^2 > 0$ as in Theorem 1, and

$$\|F_N - \Phi\|_\infty = O(N^{-1/2}),$$

where the $O(\cdot)$ is uniform over all folds. Here, Φ denotes the distribution function of the normal distribution with mean 0 and variance σ^2.

Proof: The central limit theorem for fold energies follows by applying a result on m-*dependent random fields* given in [20]. Slightly specifying to our case and using our notation, it can be stated as follows.

Theorem 4. *Suppose that for some $M > 0$, $\mathbb{E}[X_i^8] \leq M < \infty$ for all i and that $\{X_i(f|S)\}_{i\in[N]}$ is m-dependent, for some $m > 0$. Provided $\liminf_{N\to\infty} \frac{\mathbf{Var}\ [\mathcal{E}(f,S)]}{N} > 0$, we have*

$$\|F_N - \Phi\|_\infty = O(N^{-1/2}).$$

We first establish m-dependence. This follows easily from the fact that the local energy of a node i in a given fold can only be dependent on the local energies of those nodes j that are within a lattice-adjacency neighborhood of i of some fixed, finite radius. This, in turn, follows from the independent choice of the sequence elements. Thus, we have m-dependence with $m = 2$.

It is further required that the variance of $\mathcal{E}(f, S)$ grows at least linearly with N. We shall establish that $\mathbf{Var}\ [\mathcal{E}(f, S)] = \Theta(N)$ (for a large class of Q), and, along the way, derive the polynomial $V(p)$ whose existence is claimed in the lemma statement. We have

$$\mathbf{Var}\ [\mathcal{E}(f, S)] = \sum_{i=1}^{N} \mathbf{Var}\ [X_i] + 2 \sum_{1\leq i<j\leq N} \mathrm{Cov}[X_i, X_j].$$

Since $N-o(N)$ nodes have exactly two contacts, the dominant contribution to the first sum comes from those nodes, all of which have the same variance $v(p)$, a polynomial in p with coefficients that are polynomials in the entries of Q.

Note, then, that if nodes i and j are not lattice-adjacent, then $\mathrm{Cov}[X_i, X_j] = 0$. Thus, any node i is involved in at most 3 nonzero covariance terms. In fact, $N - o(N)$ nodes are involved in exactly 2 such terms. All such nodes i and j have covariance equal to some fixed $r(p)$, a polynomial in p with coefficients that are polynomials in the variables Q_{HH}, Q_{HP}, Q_{PP}.

By conditioning on the symbols assigned to nodes i and j and their other two lattice neighbors, both $v(p)$ and $r(p)$ can be computed exactly. Thus, we have $\mathbf{Var}\ [\mathcal{E}(f, S)] = N \cdot (v(p) + 2r(p)) + o(N)$. We call $V(p) = v(p) + 2r(p)$ the *variance polynomial* of Q. Provided it is not identically 0, it has finitely many roots, at which the variance of each fold energy is $o(N)$. Excluding these roots, the variance is $\Theta(N)$, as claimed, and we set $\sigma^2 = V(p)$.

Finally, it is required that, for all i, $\mathbb{E}[X_i^8] < \infty$. Since X_i is bounded between two constants with probability 1, all moments exist, and the proof is complete. ∎

Next, we need a lemma bounding the probability of large deviations for $\mathcal{E}(f, S)$.

Lemma 2 (Large deviations of $\mathcal{E}(f, S)$)**.** *There exists a constant $C > 0$ such that, for any $t > 0$ and $f \in \mathcal{F}_N$,*

$$\Pr[|\mathcal{E}(f, S) - \mathbb{E}[\mathcal{E}(f, S)]| \geq tN] \leq 2\exp\left(-\frac{t^2 N}{C}\right).$$

Proof: The proof uses the fact that each node energy is dependent on at most a constant number of others to bound the martingale differences.

To be precise, we define the filtration $(\mathbb{F}_i)_{i=0}^N$ by

$$\mathbb{F}_i = \sigma(X_1(f|S), \ldots, X_i(f|S)),$$

and then we define $(Y_i)_{i=0}^N$ to be the Doob martingale of $\mathcal{E}(f, S)$ with respect to $(\mathbb{F}_i)$, that is, $Y_i = \mathbb{E}[\mathcal{E}(f, S)|\mathbb{F}_i]$. To apply Hoeffding's inequality, we need to show that the martingale differences are bounded:

$$\begin{aligned}|Y_i - Y_{i-1}| = |&\mathbb{E}[X_1(f|S) + \cdots + X_N(f|S)|\mathbb{F}_i] \\ &- \mathbb{E}[X_1(f|S) + \cdots + X_N(f|S)|\mathbb{F}_{i-1}]|.\end{aligned}$$

Now, we partition the terms comprising the expectation defining Y_i into those which are dependent on $X_i(f|S)$ and those which are not: we define $A = \{j|X_j(f|S) \perp X_i(f|S)\}$, and then we note that, for any $j \in A$,

$$\mathbb{E}[X_j(f|S)|\mathbb{F}_i] = \mathbb{E}[X_j(f|S)|\mathbb{F}_{i-1}].$$

Thus, those terms whose indices are in A cancel in the expression for $|Y_i - Y_{i-1}|$, leaving

$$|Y_i - Y_{i-1}| = |\sum_{j\notin A} (\mathbb{E}[X_j(f|S)|\mathbb{F}_i] - \mathbb{E}[X_j(f|S)|\mathbb{F}_{i-1}])|.$$

All local energies are bounded above by some fixed constant, and, by the m-dependence property of the local energies, $|A|$ is also bounded above by a fixed constant. Thus, there is some fixed L such that, for all $f \in \mathcal{F}_N$ and $i \in [N]$, $|Y_i - Y_{i-1}| \leq L$. Applying Hoeffding's inequality with this bound then yields the claimed result. ∎

Lemmas 1 and 2 are then sufficient to derive an estimate of the MGF of a normalized fold energy.

Lemma 3 (Asymptotics of the MGF of $\hat{\mathcal{E}}_N$)**.** *Let $\phi_N : \mathbb{R} \to \mathbb{R}$ denote the MGF of a generic normalized fold energy:*

$$\phi_N(t) = \mathbb{E}[e^{t\frac{\mathcal{E}(f,S)-\mathbb{E}[\mathcal{E}(f,S)]}{\sqrt{N}}}].$$

We have, for arbitrary fixed $t \in \mathbb{R}$,

$$\begin{aligned}\phi_N(t\sqrt{N}) &= e^{N\log\phi(t)}(1 + O(N^{-1/2})) \\ &= e^{\frac{1}{2}\sigma^2 t^2 N}(1 + O(N^{-1/2})).\end{aligned}$$

Here, $\phi(t)$ denotes the MGF of the normal distribution with mean 0 and variance σ^2.

Proof: The strategy is to show that the tails of the integral defining $\phi_N(t\sqrt{N})$ are negligible, leaving a central region that can be handled via Lemma 1.

We first handle the degenerate case of $t = 0$. In this case, $\phi_N(t\sqrt{N}) = \phi_N(0) = \mathbb{E}[e^0] = 1$ and the claim holds.

We now move on to the case where $t > 0$. Let $F_N(x)$ be the distribution function of $\hat{\mathcal{E}}_N$ (recall that this is the centered and normalized energy). Then $\phi_N(t\sqrt{N})$ is given by

$$\phi_N(t\sqrt{N}) = \int_{-\infty}^{\infty} e^{t\sqrt{N}x}\, \mathrm{d}F_N(x).$$

Taking the tail at $\theta\sqrt{N}$ of this integral, for some θ which we will choose later, yields $\int_{\theta\sqrt{N}}^{\infty} e^{t\sqrt{N}x}\,\mathrm{d}F_N(x)$. Defining $g(x) = e^{t\sqrt{N}x}$ for brevity, we evaluate the above integral by parts:

$$\begin{aligned}
&e^{t\sqrt{N}b}F_N(b) - e^{t\theta N}F_N(\theta\sqrt{N}) - \int_{\theta\sqrt{N}}^{b} F_N(x)\,\mathrm{d}g(x)\\
&= e^{t\sqrt{N}b}(1-F_N(b)) + e^{t\theta N}(1-F_N(\theta\sqrt{N}))\\
&+ \int_{\theta\sqrt{N}}^{b}(1-F_N(x))t\sqrt{N}e^{t\sqrt{N}x}\,\mathrm{d}x,
\end{aligned}$$

where the equality is by adding and subtracting 1 inside the integral. Upper bounding using Lemma 2 gives

$$2e^{t\sqrt{N}b-\frac{b^2}{C}} + 2e^{t\theta N-\frac{\theta^2 N}{C}} + \int_{\theta\sqrt{N}}^{b}(1-F_N(x))t\sqrt{N}e^{t\sqrt{N}x}\,\mathrm{d}x.$$

Taking $b\to\infty$, the first term tends to 0, and the upper limit on the integral becomes ∞. As for the second term, we observe that

$$e^{t\theta N-\theta^2 N/C}e^{-N(\theta^2/C-t\theta)}.$$

Thus, if we choose θ to satisfy

$$\theta^2/C - t\theta > 0 \iff \theta > Ct,$$

the second term is $o(1)$ as $N\to\infty$.

It remains to bound the contribution of the integral. We again apply Lemma 2, which gives

$$\int_{\theta\sqrt{N}}^{\infty}(1-F_N(x))t\sqrt{N}e^{t\sqrt{N}x}\,\mathrm{d}x \le 2t\sqrt{N}\int_{\theta\sqrt{N}}^{\infty} e^{-x^2/C+t\sqrt{N}x}\,\mathrm{d}x.$$

Now, we write the exponent inside the integral as

$$-x^2/C + t\sqrt{N}x = -x^2(1/C - t\sqrt{N}/x).$$

Since $x\ge\theta\sqrt{N}$, the expression inside the parentheses is at least some positive constant L, since $\theta > Ct$. It is not hard to see that the integral is then $\Theta(e^{-\theta^2 N})$, so that the entire expression is $o(1)$ as $N\to\infty$.

The other tail of the MGF integral is easily handled:

$$\begin{aligned}
&\mathbb{E}[e^{t\sqrt{N}\hat{\mathcal{E}}(f,S)}\mathbb{I}[\hat{\mathcal{E}}(f,S)\le -\theta\sqrt{N}]]\\
&\le e^{-t\theta N}\Pr[\hat{\mathcal{E}}(f,S)\le -\theta\sqrt{N}]\\
&\le e^{-t\theta N} = o(1).
\end{aligned}$$

In the case where $t<0$, we switch the tails in the above bounds.

This leaves the central region (for any t):

$$\begin{aligned}
\int_{-\theta\sqrt{N}}^{\theta\sqrt{N}} e^{t\sqrt{N}x}\,\mathrm{d}F_N(x) &= (1+O(N^{-1/2}))\int_{-\theta\sqrt{N}}^{\theta\sqrt{N}} e^{t\sqrt{N}x}\,\mathrm{d}\Phi(x)\\
&= (1+O(N^{-1/2}))\int_{-\infty}^{\infty} e^{t\sqrt{N}x}\,\mathrm{d}\Phi(x)\\
&= (1+O(N^{-1/2}))e^{\frac{1}{2}t^2\sigma^2 N}.
\end{aligned}$$

Here, the first equality is by Lemma 1 (a more detailed explanation will follow), and the asymptotic equivalence follows from the fact that the tails of the Gaussian distribution are negligible. To be more precise, we first observe that we can ignore the lower half of the integral. In the case where $t>0$, we have

$$\begin{aligned}
\int_{-\theta\sqrt{N}}^{0} e^{t\sqrt{N}x}\,\mathrm{d}F(x) &\le \int_{-\infty}^{0} e^{t\sqrt{N}x}\,\mathrm{d}F(x)\\
&\le e^{t\sqrt{N}0}\int_{-\infty}^{0}\mathrm{d}F(x) \le 1 = \Theta(1),
\end{aligned}$$

which is negligible. Now, applying integration by parts to the remaining integral, we have

$$\begin{aligned}
&\int_{0}^{\theta\sqrt{N}} e^{t\sqrt{N}x}\,\mathrm{d}F_N(x)\\
&= e^{tN\theta}F_N(\theta\sqrt{N}) - e^0 F_N(0) - \int_0^{\theta\sqrt{N}} F_N(x)\,\mathrm{d}e^{t\sqrt{N}x}.
\end{aligned}$$

According to Lemma 1,

$$F_N(x) = \Phi(x) + O(N^{-1/2}),$$

where $\Phi(x)$ is the cumulative distribution function of the normal distribution with mean 0 and variance σ^2, and the $O(\cdot)$ is uniform with respect to x. Since, in the range under consideration, $x\ge 0$, $\Phi(x)\in[1/2,1)$, so that this implies $F_N(x) = \Phi(x)(1+O(N^{-1/2}))$. Substituting this into the expression for the integral, we get

$$(1+O(N^{-1/2}))[e^{tN\theta}\Phi(\theta\sqrt{N}) - e^0\Phi(0) - \int_0^{\theta\sqrt{N}}\Phi(x)\,\mathrm{d}e^{t\sqrt{N}x}],$$

and applying the integration by parts formula again yields

$$\begin{aligned}
&(1+O(N^{-1/2}))\int_0^{\theta\sqrt{N}} e^{t\sqrt{N}x}\,\mathrm{d}\Phi(x)\\
&= (1+O(N^{-1/2}))\int_{-\infty}^{\infty} e^{t\sqrt{N}x}\,\mathrm{d}\Phi(x) + O(1)\\
&= (1+O(N^{-1/2}))\int_{-\infty}^{\infty} e^{t\sqrt{N}x}\,\mathrm{d}\Phi(x).
\end{aligned}$$

where the integral is precisely the moment generating function of $\mathcal{N}(0,\sigma^2)$, evaluated at $t\sqrt{N}$. The added term $O(1)$ comes from completing the lower tail. The case where $t\le 0$ is handled similarly. Finally, taking a logarithm, dividing by N, and taking $N\to\infty$ gives the desired expression. ∎

Using the expression developed in Lemma 3, we can finally begin the derivation of the claimed free energy bounds. For the first upper bound,

$$\begin{aligned}
&\mathbb{E}[\log Z(S,\beta)] \le \log\mathbb{E}[Z(S,\beta)]\\
&= \log\sum_{f\in\mathcal{F}_N} e^{-\beta\mathbb{E}[\mathcal{E}(f,S)]}\mathbb{E}\left[e^{-\beta\sqrt{N}\frac{\mathcal{E}(f,S)-\mathbb{E}[\mathcal{E}(f,S)]}{\sqrt{N}}}\right]\\
&= \log\sum_{f\in\mathcal{F}_N} e^{-\beta\mathbb{E}[\mathcal{E}(f,S)]}\mathbb{E}\left[e^{-\beta\sqrt{N}\hat{\mathcal{E}}_N}\right]\\
&= \log\sum_{f\in\mathcal{F}_N} e^{-\beta\alpha N(1+O(N^{-1/2}))}\cdot e^{\frac{1}{2}\sigma^2\beta^2 N}(1+O(N^{-1/2}))\\
&= \log|\mathcal{F}_N| - \beta\alpha N(1+O(N^{-1/2})) + \frac{1}{2}\sigma^2\beta^2 N + O(N^{-1/2}),
\end{aligned}$$

where we used Jensen's inequality to bring the expectation into the logarithm, and we used the fact that all of the relative errors are uniform over the set of folds. We thus have

$$\gamma(\beta) \leq 1 - \beta\alpha/\mu + \frac{1}{2}\sigma^2\beta^2/\mu,$$

and the claimed inequality (7) follows.

For the second upper bound, the strategy is to find an upper bound on the derivative with respect to β of the function $\phi(\beta) = \mathbb{E}[\log Z(S,\beta)]$.

We have

$$\begin{aligned}
&-\beta \min_{f\in\mathcal{F}_N} \mathcal{E}(f,S) \leq \log\left(\sum_{f\in\mathcal{F}_N} e^{-\beta\mathcal{E}(f,S)}\right)\\
&\implies \mathbb{E}[-\min_{f\in\mathcal{F}_N} \mathcal{E}(f,S)]\\
&\leq \beta^{-1}\log|\mathcal{F}_N| - \alpha N(1+O(N^{-1/2}))\\
&+\frac{1}{2}\sigma^2\beta N + O(\beta^{-1}N^{-1/2}),
\end{aligned}$$

where the first inequality is elementary, and the second is due to the first upper bound. We find that setting $\beta = \beta_* = \frac{\sqrt{2\mu_N}}{\sigma}$ minimizes the upper bound, yielding

$$\mathbb{E}[-\min_{f\in\mathcal{F}_N} \mathcal{E}(f,S)] \leq \sqrt{2\sigma^2\mu_N}N - \alpha N + O(\sqrt{N}).$$

Furthermore, for arbitrary β,

$$\begin{aligned}
\phi'(\beta) &= \mathbb{E}\left[-\frac{\sum_{f\in\mathcal{F}_N}\mathcal{E}(f,S)e^{-\beta\mathcal{E}(f,S)}}{\sum_{f\in\mathcal{F}_N}e^{-\beta\mathcal{E}(f,S)}}\right]\\
&\leq \mathbb{E}\left[\left(-\min_{f\in\mathcal{F}_N}\mathcal{E}(f,S)\right)\frac{Z(S,\beta)}{Z(S,\beta)}\right]\\
&= \mathbb{E}[-\min_{f\in\mathcal{F}_N}\mathcal{E}(f,S)].
\end{aligned}$$

Now, for $\beta > \beta_*$, $\phi(\beta) \leq \phi(\beta_*) + \phi'(\beta_*)(\beta-\beta_*)$, since $\phi(\beta)$ is known to be convex (a consequence of Hölder's inequality). Applying the upper bounds for $\phi'(\beta_*)$ and for $\phi(\beta_*)$ yields the second upper bound in the theorem:

$$\phi(\beta) = \mathbb{E}[\log Z(S,\beta)] \leq \beta N(\sqrt{2\sigma^2\mu_N} - \alpha + O(N^{-1/2})).$$

B. Proof of Theorem 2

The key idea here is to choose the scoring matrix Q so as to minimize the covariance between the energies of any two contacts that share only one node. For such a matrix, we then derive an explicit asymptotic formula for the variance of the partition function in terms of the square of its expected value and the MGF of the number of shared contacts between two randomly chosen folds. This MGF arises from the fact that the covariance between the energies of two folds varies linearly with the number of shared contacts.

The formula for the variance then implies, by Chebyshev's inequality, an upper bound on the probability that the partition function is much smaller than its expected value. Computing $\mathbb{E}[\log Z]$ by conditioning on this event then yields the desired result.

For the lower bound, it turns out to be beneficial to express fold energies in terms of the local energies of its *contacts*, instead of its nodes as we did in the upper bound. For a contact c (i.e., an unordered pair of distinct sequence elements) labeled by a sequence s, denote its energy by $Y_c(s)$. To aid intuition, we remark that a typical node energy (say, of node i) is expressible in terms of two contact energies: if node i makes contact with nodes j and j', then

$$X_i = Y_{\{i,j\}} + Y_{\{i,j'\}}.$$

Then the energy of a fold f is given by

$$\mathcal{E}(f,S) = 2\sum_{\text{contacts } c \text{ in } f} Y_c(S),$$

where the 2 is again from the fact that local energies are counted twice, as in (1).

For two contacts c_1, c_2 with $|c_1 \cap c_2| = 1$, we compute the covariance of the energies with respect to $S \sim \mathcal{B}_N(p)$, with p as in the theorem:

$$\begin{aligned}
&\mathrm{Cov}[Y_{c_1}(S), Y_{c_2}(S)]\\
&= p(Q_{HH}^2p^2 + 2Q_{HH}Q_{HP}p(1-p) + Q_{HP}^2(1-p)^2)\\
&+ (1-p)(Q_{PP}^2(1-p)^2 + 2Q_{HP}Q_{PP}p(1-p) + Q_{HP}^2p^2)\\
&- (p^2Q_{HH} + 2p(1-p)Q_{HP} + (1-p)^2Q_{PP})^2,
\end{aligned}$$

Defining $x = Q_{HH}, y = Q_{HP}$, and $z = Q_{PP}$, this polynomial becomes

$$\begin{aligned}
&\mathrm{Cov}[Y_{c_1}(S), Y_{c_2}(S)] = f(x,y,z)\\
&= p(x^2p^2 + 2xyp(1-p) + y^2(1-p)^2)\\
&+ (1-p)(z^2(1-p)^2 + 2yzp(1-p) + y^2p^2)\\
&- (p^2x + 2p(1-p)y + (1-p)^2z)^2.
\end{aligned}$$

and we seek a nontrivial zero. We set $y = 0$ and $z = 1$, which reduces it to

$$f(x,0,1) = (p^3-p^4)x^2 - 2p^2(1-p)^2x + ((1-p)^3 - (1-p)^4) = 0.$$

It is then easily checked that whenever $p \neq 0, 1$, there exists $x \in \mathbb{R}$ for which $f(x,0,1) = 0$. Moreover, to check that this x is such that α and σ (as in Theorem 1) are both nonzero, we note that the former is 0 only when $x = -1$. We view $f(-1,0,1)$ as a polynomial in p, and it is easy to see that this is only 0 when $p = 0$ or 1. Moreover, σ cannot be 0, since a node energy may take on two different values with positive probability.

In what follows, we assume that p and Q have been chosen so that

$$\mathrm{Cov}[Y_{c_1}(S), Y_{c_2}(S)] = 0$$

and $\alpha, \sigma \neq 0$. For any two folds $f, g \in \mathcal{F}_N$, we define $k_{f,g}$ to be the number of contacts which are in both f and g. We now relate $\mathbf{Var}\ [Z(S,\beta)]$ to $\mathbb{E}[Z(S,\beta)]^2$ with the following lemma.

Lemma 4. *We have*

$$\mathbf{Var}\ [Z(S,\beta)] = \mathbb{E}_S[Z(S,\beta)]^2(\mathbb{E}_K[e^{3\beta^2\sigma^2 K}]-1)(1+O(N^{-1/2})).$$

Proof: We calculate the second moment of $Z(S,\beta)$. Define $\tilde{\mathcal{E}}(f,S) = \mathcal{E}(f,S) - \mathbb{E}[\mathcal{E}(f,S)]$. Then

$$\begin{aligned}
&\mathbb{E}[Z(S,\beta)^2] \\
&= \sum_{f,g\in\mathcal{F}_N} \mathbb{E}_S[\exp(-\beta(\tilde{\mathcal{E}}(f,S)+\tilde{\mathcal{E}}(g,S)))]\cdot e^{2\beta\mathbb{E}[\mathcal{E}(f,S)]} \\
&= \sum_{k=0}^{N} \sum_{f,g:\ k_{f,g}=k} \mathbb{E}_S[\exp(-\beta(\tilde{\mathcal{E}}(f,S)+\tilde{\mathcal{E}}(g,S)))]e^{2\beta\mathbb{E}[\mathcal{E}(f,S)]},
\end{aligned} \tag{14}$$

simply by partitioning the set of pairs of folds into those with exactly k shared contacts, for $k = 0,\dots,N$. Next, we show that the MGFs in the expression above can be approximated by MGFs of analogous normal random variables. We claim that, for each f, g above,

$$\begin{aligned}
&\mathbb{E}_S[\exp(-\beta(\tilde{\mathcal{E}}(f,S)+\tilde{\mathcal{E}}(g,S)))] \\
&= (1+O(N^{-1/2}))\mathbb{E}[\exp(-\beta(\tilde{\mathcal{E}}_{\mathcal{N}}(f)+\tilde{\mathcal{E}}_{\mathcal{N}}(g)))],
\end{aligned} \tag{15}$$

where $\tilde{\mathcal{E}}_{\mathcal{N}}(f)$ and $\tilde{\mathcal{E}}_{\mathcal{N}}(g)$ are jointly normally distributed random variables with mean 0, variance $\sigma^2 N$, and covariance $\sigma^2 k_{f,g}$. To do this, we first calculate the covariance of $\tilde{\mathcal{E}}(f,S)$ and $\tilde{\mathcal{E}}(g,S)$ (equivalently, $\mathcal{E}(f,S)$ and $\mathcal{E}(g,S)$). Let $\mathfrak{C}(f)$ denote the set of contacts in the fold f. Then

$$\begin{aligned}
&\mathrm{Cov}[\mathcal{E}(f,S),\mathcal{E}(g,S)] \\
&= 4 \sum_{c\in\mathfrak{C}(f)\cap\mathfrak{C}(g)} \mathbf{Var}\ [Y_c(S)] \\
&+ 4 \sum_{c\in\mathfrak{C}(f)\neq c'\in\mathfrak{C}(g)} \mathrm{Cov}[Y_c(S), Y_{c'}(S)]
\end{aligned}$$

The second sum is 0, by our choice of scoring matrix, and each term of the first sum is $\sigma^2/4$, by direct calculation. Thus, we have shown that $\mathrm{Cov}[\mathcal{E}(f,S),\mathcal{E}(g,S)] = k_{f,g}\sigma^2$. Now, we follow the steps of the proof of Lemma 3. In particular, it is enough to establish a central limit theorem and a large deviations bound for the sum of the two fold energies. Both are immediate, as the two fold energies may be written as sums of node energies, and these node energies are m-dependent for some large enough constant m. Moreover, the variance of the sum is easily seen to be positive, since the covariance of the two energies is non-negative. Thus, we may conclude (15) from the proof of Lemma 3.

Continuing the derivation of $\mathbb{E}[Z(S,\beta)^2]$, we note that

$$\begin{aligned}
\tilde{\mathcal{E}}_{\mathcal{N}}(f)+\tilde{\mathcal{E}}_{\mathcal{N}}(g) &\overset{D}{=} (U(k)+W_1(k))+(U(k)+W_2(k)) \\
&= 2U(k)+W_1(k)+W_2(k),
\end{aligned}$$

where $U(k), W_1(k)$, and $W_2(k)$ are all independent, with $U(k)\sim\mathcal{N}(0,\sigma^2 k)$ and $W_1(k), W_2(k)\sim\mathcal{N}(0,\sigma^2(N-k))$. By this representation and the independence of $U(k)$, $W_1(k)$, and $W_2(k)$, we then have that (14) is equal to

$$\begin{aligned}
&(1+O(N^{-1/2}))e^{2\beta\mathbb{E}[\mathcal{E}(f,S)]} \\
&\sum_{k=0}^{N} \sum_{f,g:\ k_{f,g}=k} \mathbb{E}[e^{-2\beta U(k)}]\mathbb{E}[e^{-\beta W_1(k)}]^2.
\end{aligned}$$

To bring $\tilde{\mathcal{E}}_{\mathcal{N}}(f)$ back into the formula, we add and subtract an independent copy $\tilde{U}(k)\overset{D}{=}U(k)$ in the exponent of the second factor of the sum:

$$\begin{aligned}
\mathbb{E}[e^{-\beta W_1(k)}] &= \mathbb{E}[e^{-\beta(W_1(k)+\tilde{U}(k))}]\mathbb{E}[e^{\beta U(k)}] &(16)\\
&= \mathbb{E}[e^{-\beta\tilde{\mathcal{E}}_{\mathcal{N}}(f)}]\mathbb{E}[e^{\beta U(k)}], &(17)
\end{aligned}$$

where the first equality is by independence and equality of distribution between $U(k)$ and $\tilde{U}(k)$, and the second is by the fact that $W_1(k)+\tilde{U}(k)\overset{D}{=}\tilde{\mathcal{E}}_{\mathcal{N}}(f)$. We then pull the first factor of (17) out of the sums (since it is the same for all f), and this leaves

$$\begin{aligned}
&(1+O(N^{-1/2}))\mathbb{E}[e^{-\beta(\tilde{\mathcal{E}}_{\mathcal{N}}(f)+\mathbb{E}[\mathcal{E}(f,S)])}]^2 \\
&\cdot \sum_{k} \sum_{f,g:\ k_{f,g}=k} \mathbb{E}[e^{-2\beta U(k)}]\mathbb{E}[e^{\beta U(k)}]^2.
\end{aligned}$$

Now, the terms of the inner sum are independent of f and g, so that the outer sum becomes

$$\begin{aligned}
&|\mathcal{F}_N|^2 \sum_{k} \frac{\#\{f,g\in\mathcal{F}_N:\ k_{f,g}=k\}}{|\mathcal{F}_N|^2}\mathbb{E}[e^{-2\beta U(k)}]\mathbb{E}[e^{\beta U(k)}]^2 \\
&= |\mathcal{F}_N|^2\mathbb{E}_K[\mathbb{E}[e^{-2\beta U(K)}]\mathbb{E}[e^{\beta U(K)}]^2].
\end{aligned}$$

Since $U(k)$ is normally distributed, we can compute its MGF, and this reduces the formula above to

$$|\mathcal{F}_N|^2\mathbb{E}_K[e^{3\beta^2\sigma^2 K}].$$

Now, looking at the factors of the entire expression outside of the expectation with respect to K,

$$\begin{aligned}
&|\mathcal{F}_N|^2\mathbb{E}[e^{-\beta(\tilde{\mathcal{E}}_{\mathcal{N}}(f)+\mathbb{E}[\mathcal{E}(f,S)])}]^2 \\
&= (1+O(N^{-1/2}))|\mathcal{F}_N|^2\mathbb{E}[e^{-\beta(\tilde{\mathcal{E}}(f,S)+\mathbb{E}[\mathcal{E}(f,S)])}]^2 \\
&= (1+O(N^{-1/2}))|\mathcal{F}_N|^2\mathbb{E}[e^{-\beta\mathcal{E}(f,S)}]^2 \\
&= (1+O(N^{-1/2}))\mathbb{E}[Z(S,\beta)]^2.
\end{aligned}$$

Here, the first equality is by the proof of Lemma 3, the second is by definition of $\tilde{\mathcal{E}}(f,S)$, and the third is by linearity of expectation and the definition of the partition function. This completes the proof. ∎

Given Lemma 4, we now prove the claimed lower bound of Theorem 2. We define the event

$$A = A_\epsilon = [Z \geq \epsilon\mathbb{E}[Z]]$$

for arbitrary $\epsilon > 0$. Then Chebyshev's inequality gives

$$1-\Pr[A] \leq \Pr[|Z-\mathbb{E}[Z]| \geq (1-\epsilon)\mathbb{E}[Z]] \leq \frac{\mathbf{Var}\ [Z]}{(1-\epsilon)^2\mathbb{E}[Z]^2}.$$

By Lemma 4, this becomes

$$1 - \Pr[A] \le \frac{\mathbb{E}_K[e^{3\beta^2\sigma^2 K}] - 1}{(1-\epsilon)^2}.$$

In other words,

$$\Pr[A] \ge 1 - \frac{\mathbb{E}_K[e^{3\beta^2\sigma^2 K}] - 1]}{(1-\epsilon)^2},$$

and we denote this lower bound by p_A. We can choose $\epsilon \xrightarrow{N\to\infty} 1^-$ sufficiently slowly so that $p_A = 1 - o(1)$ (e.g., $\epsilon = 1-(\mathbb{E}_K[e^{3\beta^2\sigma^2 K}]-1)^{(1-\delta)/2}$, for a small positive constant δ). Then

$$\mathbb{E}[\log Z] = \mathbb{E}[\log Z|A]\Pr[A] + \mathbb{E}[\log ZI[\neg A]] \tag{18}$$
$$\ge (\log \mathbb{E}[Z] + \log\epsilon)\Pr[A] + \mathbb{E}[\log ZI[\neg A]]. \tag{19}$$

First term: We can explicitly compute $\log \mathbb{E}[Z]$ as

$$\begin{aligned}
&\log(|\mathcal{F}_N|\mathbb{E}[e^{-\beta\mathcal{E}(f,S)}]) \\
&= \log|\mathcal{F}_N| + \log e^{-\beta\alpha N + \frac{1}{2}\beta^2\sigma^2 N} + O(N^{-1/2}) \\
&= \log|\mathcal{F}_N| - \beta\alpha N + \frac{1}{2}\beta^2\sigma^2 N + O(N^{-1/2}),
\end{aligned}$$

where we applied Lemma 3 to estimate the MGF. Thus, the first term is lower bounded by

$$(\log|\mathcal{F}_N| - \beta\alpha N + \frac{1}{2}\beta^2\sigma^2 N + o(1))p_A.$$

Here, the $o(1)$ comes from $\log\epsilon$, recalling that we chose ϵ so that $\epsilon = 1 - o(1)$.

Second term: Since $I[\neg A] \ge 0$, we choose an arbitrary $f \in \mathcal{F}_N$ (we may be able to refine this to produce a better bound), and then

$$\log Z \ge \log e^{-\beta\mathcal{E}(f,S)} = -\beta\mathcal{E}(f,S).$$

Then

$$\begin{aligned}
&\mathbb{E}[\log ZI[\neg A]] \\
&\ge \mathbb{E}[-\beta\mathcal{E}(f,S)I[\neg A]] \\
&= -\beta\mathbb{E}[\tilde{\mathcal{E}}(f,S)I[\neg A]] - \beta\mathbb{E}[\mathcal{E}(f,S)]\Pr[\neg A] \\
&\ge -\beta\mathbb{E}[|\tilde{\mathcal{E}}(f,S)|I[\neg A]] - \beta\mathbb{E}[\mathcal{E}(f,S)]\Pr[\neg A] \\
&\ge -\beta\mathbb{E}[|\tilde{\mathcal{E}}(f,S)|] - \beta\mathbb{E}[\mathcal{E}(f,S)]\Pr[\neg A]. \qquad (20)
\end{aligned}$$

Because $\tilde{\mathcal{E}}(f,S) \sim \mathcal{N}(0,\Theta(N))$, the first term of (20) is $\Theta(\sqrt{N})$. The second term of (20) is $\Theta(N)\Pr[\neg A]$, which is upper bounded by

$$\Theta(N)(1-p_A),$$

which, by the hypothesis (9) on the MGF of K, is $o(N)$.

Putting everything together: We thus have a lower bound on the free energy given by

$$\frac{\mathbb{E}[\log Z]}{N} \ge p_A(\mu_N - \beta\alpha + \frac{1}{2}\beta^2\sigma^2 + o(N^{-1})) + o(1).$$

C. Proof of Corollary 1

The claim follows from the representation (4) and the lower bound on $\mathbb{E}[\log Z]$ given in Theorem 2. This gives a lower bound of

$$\begin{aligned}
&H(F|S) \\
&\ge (1-o(1))(\log|\mathcal{F}_N| - \beta N(\alpha - \alpha_*(\beta)) + \frac{1}{2}\beta^2\sigma^2 N + o(1)) \\
&- \beta\Theta(N)(1-p_A).
\end{aligned}$$

Since $\beta N(\alpha - \alpha_*(\beta)) = \beta N\psi(N) = o(1)$ and $\beta^2 N = o(N^{-4/3+1}) = o(N^{-1/3})$, the first term of the lower bound is

$$\log|\mathcal{F}_N| - o(1).$$

Meanwhile, to estimate the second term, since $\beta = o(N^{-2/3})$ and $K \le N + O(\sqrt{N})$,

$$\mathbb{E}_K[e^{c\beta^2 K}] = \mathbb{E}_K[1 + c\beta^2 K + O(\beta^4 N^2)] \le 1 + \beta^2\Theta(N).$$

Then

$$1 - p_A \le \beta^2\Theta(N),$$

so that the second term is upper bounded in absolute value by

$$\beta^3\Theta(N^2).$$

Since $\beta = o(N^{-2/3})$, this is $o(1)$, and (13) is verified.

D. Proof of Theorem 3

First, we show that for any sequence s and fold f of length N,

$$\mathcal{E}(f,s) = -D_s(H) + O(\sqrt{N}), \tag{21}$$

where $D_s(H)$ denotes the number of Hs in the string s. To do this, we consider the *contact graph* of an arbitrary fold f (we denote it by $G(f)$), which we define as follows: the vertices are all of the nodes of the walk, except for the endpoints (this is for simplicity). There is an edge between two vertices if and only if they form a contact. We observe a few structural characteristics of this graph: it has $N - O(\sqrt{N})$ nodes with degree exactly 2, since, for any node x in f which is neither an endpoint nor on the boundary (which has size $O(\sqrt{N})$), x has exactly two sequence neighbors, leaving exactly two contact neighbors. Moreover, $G(f)$ contains no cycles, since a cycle would imply that f is not connected. Thus, the connected components of $G(f)$ are string graphs, and s induces a labeling on them. Because at most $O(\sqrt{N})$ nodes have degree 1, there are at most $O(\sqrt{N})$ components.

We consider the contribution to $\mathcal{E}(f,s)$ of the labeling X of an arbitrary component of f. We claim that

$$\mathcal{E}(X,s) = -D_X(H) + O(1),$$

after which summing over all components will give (21). To compute the energy of X, we divide it into chunks of Hs: a *boundary* chunk is of the form H^k (if $X = H^k$), H^kP,

or PH^kP, and there are only at most two of them. In all such cases, the sum of contributions of contacts between Hs (recall that all contacts are counted twice) is $-(k-1)$, and the contact between H and P contributes $O(1)$, so that the total contribution of a boundary chunk is $-k+O(1)$.

For a *non-boundary* chunk, which is of the form PH^kP, the contribution of contacts between Hs is again $-k+1$, while the PH/HP contacts contribute a total of -1, resulting in a score of $-k$. Summing over all chunks gives the claimed energy, and this establishes (21).

Now, with (21) in hand, we can compute $\mathbb{E}[\log Z(S,\beta)]$ for a random S:

$$\begin{aligned}\mathbb{E}[\log Z(S,\beta)] &= \mathbb{E}\left[\log\left(\sum_{f\in\mathcal{F}_N}\exp(-\beta\mathcal{E}(f,S))\right)\right]\\ &= \mathbb{E}[\log\left(|\mathcal{F}_N|\exp(\beta D_S(H)+O(\sqrt{N}))\right)]\\ &= \mathbb{E}[\log|\mathcal{F}_N|+\beta D_S(H)+O(\sqrt{N})]\\ &= \log|\mathcal{F}_N|+\beta\mathbb{E}[D_S(H)]+O(\sqrt{N}).\end{aligned}$$

Dividing by $\log|\mathcal{F}_N|$ and taking $N\to\infty$ gives the claimed free energy. Note that, in the special case where $S\sim\mathcal{B}_N(p)$, this becomes

$$\gamma(\beta)=1-\beta\alpha/\mu,$$

which completes the proof.

We remark that a similar calculation can be done for any scalar multiple of the chosen scoring matrix or for the matrix with the roles of P and H swapped.

REFERENCES

[1] W. Bialek, *Biophysics: Searching for Principles*. Princeton, NJ, USA: Princeton University Press, 2012.

[2] A. Magner, W. Szpankowski, and D. Kihara, "On the origin of protein superfamilies and superfolds," *Scientific Reports*, vol. 5, no. 8166, 2015.

[3] C. E. Soteros and S. G. Whittington, "Contacts in self-avoiding walks and polygons," *Journal of Physics A: Mathematical and General*, vol. 34, no. 19, 2001.

[4] M. Talagrand, *Spin Glasses: A Challenge for Mathematicians*. New York, NY, USA: Springer, 2003.

[5] N. Madras and G. Slade, *The Self-Avoiding Walk*. Birkhäuser Basel, 2013.

[6] A. Sali, E. Shakhnovich, and M. Karplus, "How does a protein fold?" *Nature*, vol. 369, no. 6477, pp. 248–251, May 1994. [Online]. Available: http://dx.doi.org/10.1038/369248a0

[7] H. K. Nakamura and M. Sasai, "Population analyses of kinetic partitioning in protein folding," *Proteins Structure, Function, and Genetics*, vol. 43, pp. 280–291, 2001.

[8] W. Szpankowski, *Average Case Analysis of Algorithms on Sequences*. New York, NY, USA: John Wiley & Sons, Inc., 2001.

[9] H. Duminil-Copin and S. Smirnov, "The connective constant of the honeycomb lattice equals $\sqrt{2+\sqrt{2}}$," *Annals of Mathematics*, vol. 175, pp. 1653–1665, 2012.

[10] H. L. Abbott and D. Hanson, "A lattice path problem," *Ars Combinatoria*, vol. 6, pp. 163–178, 1978.

[11] M. Bousquet-Mélou, A. J. Guttmann, and I. Jensen, "Self-avoiding walks crossing a square," *Journal of Physics A: Mathematical and General*, vol. 38, no. 42, 2005.

[12] M. Mézard and A. Montanari, *Information, Physics, and Computation*. New York, NY, USA: Oxford University Press, Inc., 2009.

[13] D. Sherrington and S. Kirkpatrick, "Solvable model of a spin glass," *Physical Review Letters*, vol. 35, pp. 1792–1796, 1975.

[14] G. Parisi, "A sequence of approximate solutions to the s-k model for spin glasses," *Journal of Physics A*, vol. 13, pp. L–115, 1980.

[15] M. Talagrand, "The Parisi formula," *Annals of Mathematics*, vol. 163, pp. 221–263, 2006.

[16] T. Cover and J. A. Thomas, *Elements of Information Theory*. Wiley, 2006.

[17] M. Bayati, D. Gamarnik, and P. Tetali, "Combinatorial approach to the interpolation method and scaling limits in sparse random graphs," *Annals of Probability*, vol. 41, no. 6, pp. 4080–4115, 2013.

[18] N. Kistler, "Derrida's random energy models: From spin glasses to the extremes of correlated random fields," *pre-print*, 2014. [Online]. Available: http://arxiv.org/abs/1412.0958

[19] J. M. Buhmann, A. Gronskiy, and W. Szpankowski, "Free energy rates for a class of very noisy optimization problems," *Proceedings of DMTCS, AofA, Paris*, pp. 67–78, 2014.

[20] H. Takahata, "On the rates in the central limit theorem for weakly dependent random fields," *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete*, vol. 64, 1983.