Heteropolymer sequence design and preferential solvation of hydrophilic monomers: Application of random energy model

Longhua Hu and Alexander Y. Grosberg

Department of Physics, University of Minnesota, 116 Church Street SE, Minneapolis, Minnesota 55455, USA

We study the role of the surface of the globule and the role of interactions with the solvent for designed-sequence heteropolymers using the random energy model. We investigate the ground-state energy and surface-monomer composition distribution. By comparing the freezing transition in random and designed-sequence heteropolymers, we discuss the effects of design. Based on our results, we are able to show under which conditions the solvation effect improves the quality of the sequence design. Finally, we study sequence-space entropy and discuss the number of available sequences as a function of the imposed requirements for design quality.

PACS number(s): 87.15.−v

I. INTRODUCTION

The simplest concept taught to students about protein structure is that hydrophobic monomers are mostly inside water-soluble globules, while hydrophilic monomers are mostly on the surface. This beautiful idea has been around for over half a century [1,2], everyone agrees that it represents a cornerstone for our understanding of proteins [3], and yet it is somehow neglected in the most sophisticated theories of heteropolymers with quenched sequences. Here, we have in mind the train of works which started from pioneering contributions by Bryngelson and Wolynes [4] and Shakhnovich and Gutin [5]. The insight of the former authors [4] was to recognize the deep connection between protein field and that of spin glasses and to apply the random energy model (REM) developed by Derrida [6]; the contribution of the latter [5] was to actually derive the REM approximation for a consistent microscopic model of a heteropolymer with independent random interactions. By now, it is understood that the REM is a well-controlled mean-field approximation for large compact heteropolymers [7]. The important part of heteropolymer theory was also the idea of sequence design, which was used both to better model proteins and to test heteropolymer properties in general [8].

What is important to emphasize is that heteropolymer freezing and sequence design theories operate within the so-called volume approximation, neglecting surface terms in energy. Our goal in the present work is to investigate, for the simplest tractable model, the interplay of heteropolymer freezing and sequence design with preferential solvation of some monomer species on the surface of the globule. In fact, even for random sequences preferential solvation was not included in REM-based heteropolymer theory until very recently [9]. Our work is ideologically a sequel to the paper Ref. [9], and we will use the ideas of that work.

In a recent series of works [10,11], sequence design was discussed (under the different name of "coloring") in a slightly different prospective, with an eye on chemically preparing proteinlike copolymers. The solvation effect was given a very prominent role in these works. One of the goals of our work is to make a closer link between various implementations of the sequence design paradigm.

This paper is organized as follows. First we will introduce the solvation model (Sec. II A and Fig. 1). Then we will talk about the sequence design technique and how it is affected by the solvation (Sec. II B). The surface-monomer composition distribution is obtained in Sec. III A. Our major results are summarized in the phase diagram of the system, sketched in Fig. 3. Finally, we discuss in Sec. V the availability of sequences as a function of their quality characterized by the energy gap between their ground state and the majority of other states.

II. MODEL

A. Energy: Bulk and surface terms

In our model, each monomer is assigned a quenched random variable σ, which represents its monomer type. For the random sequence, we assume that σ for each monomer is drawn from some probability distribution $p(\sigma)$. For simplicity, we restrict $p(\sigma)$ to have zero average $\int \sigma p(\sigma) d\sigma = 0$ and unit variance $\int \sigma^2 p(\sigma) d\sigma = 1$. There are 20 possible values of σ for natural proteins since the number of amino acids is 20. Theoretically it is convenient to consider a continuous distri-
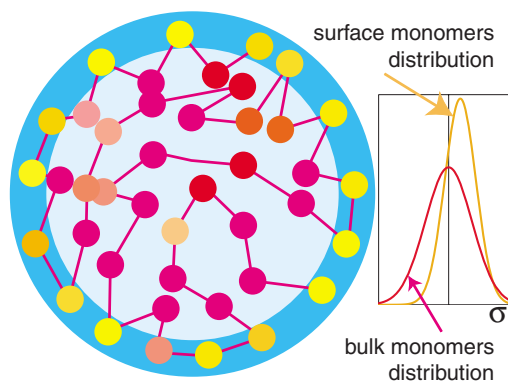

FIG. 1. (Color online) Illustration of the model. Monomers are connected by covalent bonds, and monomer type is presented by the shade. Surface is enriched with hydrophilic monomers, such that the distribution of surface monomers $f(\sigma)$ is shifted compared with the bare distribution $p(\sigma)$.

bution of σ or a discrete distribution of just two monomer types.

In our model, the energy of the system consists of contributions from direct contact between monomers and of the contribution of contacts between monomers and solvent. The former contact energy has a "homopolymeric" strong average attraction part $\bar{B}$ independent of monomer type and a "heteropolymeric" contribution $-\delta B \sigma_i \sigma_j$ with amplitude δB. We assume that $\bar{B}$ is sufficiently large, such that the globule is quite dense, and the contacts with the solvent take place only on the surface of the globule. We mostly look at the case $\delta B > 0$, such that similar monomers attract each other. We assume that each contact with the solvent provides energy $-\Gamma \sigma_i$. Thus, since $\Gamma > 0$, monomers with $\sigma > 0$ are hydrophilic, while those with $\sigma < 0$ are hydrophobic. Thus, the Hamiltonian of our model depends on the sequence, presented by seq=$\{\sigma_i\}$, and conformation, specified by positions of all monomers conf=$\{\mathbf{r}_i\}$. The Hamiltonian reads

$$H(\text{seq},\text{conf}) = \sum_{i<j}^{N} (\bar{B} - \delta B \sigma_i \sigma_j)\Delta_{ij} - \Gamma \sum_{i\in G}^{K} \sigma_i. \qquad (1)$$

Here G is the set of $K \sim N^{2/3}$ monomers located on the surface and, therefore, exposed to the solvent. $\Delta_{ij} \equiv \Delta(\mathbf{r}_i - \mathbf{r}_j)$ is contact map defined as

$$\Delta_{ij} = \begin{cases} 1, & \mathbf{r}_i \text{ and } \mathbf{r}_j \text{ are nearest neighbors,} \\ 0, & \text{otherwise.} \end{cases} \qquad (2)$$

As we mentioned, we only consider the most compact conformations and we assume there are Q contacts in every conformation, so $\Sigma_{i<j}^{N} \Delta_{ij} = Q$. For a compact globule, $Q \sim N$. For simplicity, we just use N to denote the number of contacts in a conformation.

Apart from the surface term, this model is equivalent to the one considered earlier—e.g., [12]. A schematic of the model is sketched in Fig. 1. In our model, we assume that some K monomers are exposed to the solvent because they belong to the surface layer of the globule, while other N-K monomers are not at all exposed to the solvent. This is, of course, only an approximation to reality, in which solvent exposure changes continuously. But this is a good approximation as long as we deal with a dense well-formed globule.

A polymer chain with Hamiltonian (1), with quenched sequence and in statistical equilibrium in terms of conformation, will exhibit some preference of hydrophilic monomers towards the surface—as long as $\Gamma > 0$. The way to characterize this quantitatively is to address the statistics of σ values for surface-exposed monomers. Namely, we will be interested in a distribution $f(\sigma)$ of surface monomers. We expect this distribution to be different from the bare distribution of all monomers $p(\sigma)$. Qualitatively, this is illustrated in the inset of Fig. 1.

Of course, the effect of surface exposure to the solvent depends on the sequence. In general, the hydrophilic effect adds frustration to the system. Indeed, placing a certain monomer on the surface necessitates placing its sequence neighbors close to the surface, while their identity, or their σ

values, might imply energetic preference for the interior region of the globule. To address this delicate sequence dependence, we will look at designed sequences.

B. Sequence design

A random sequence can be generated by a suitable Poisson process—i.e., by the probability distribution

$$P_{\text{seq}}^{(0)} = \prod_{i=1}^{N} p(\sigma_i). \qquad (3)$$

By sequence design, we want to bias the sequence probabilities in a controlled fashion. This can be done in the following way.

The sequence design procedure starts from the choice of the target conformation which we will denote $\star$. In our consideration here, $\star$ might be any compact conformation; in other words, we ignore the difference of designability for possible target conformations [13]—simply because designability and surface exposure are two independent effects and we do not want to further complicate our work by accounting for designability. There is no doubt that designability along with surface effects must be incorporated into the complete theory. Given a target conformation $\star$, we will consider a statistical Gibbs ensemble in which conformation $\{\mathbf{r}_i\}=\star$ is quenched, while the sequence $\{\sigma_i\}$ is annealed and comes to thermodynamic equilibrium at the design temperature T_d, which is not necessarily equal to the real temperature T. More specifically, we use the canonical sequence design scheme [12], in the sense that it is generated by the canonical ensemble of annealed sequences. This results in the following probability distribution of sequences:

$$P_{\text{seq}}^{\star} = P_{\text{seq}}^{(0)} \frac{\exp[-H^d(\text{seq},\star)/T_d]}{\sum_{\text{seq}'} P_{\text{seq}'}^{(0)} \exp[-H^d(\text{seq}',\star)/T_d]}, \qquad (4)$$

where the denominator ensures normalization. Of course, this is just the scheme; the real features of the ensemble of designed sequences are controlled by the design Hamiltonian, $H^d(\text{seq},\star)$. It might be the same Hamiltonian as in Eq. (1), but this is not at all necessary. Moreover, it is useful to explore the more general situation in which $H^d \neq H$. We will use a design Hamiltonian of the same functional form as in Eq. (1), but with the different parameters δB^d and Γ^d ($\bar{B}^d$, although formally included for symmetry, does not play any role in sequence design, because this term in energy is sequence independent):

$$H^d(\text{seq},\star) = \sum_{i<j}^{N} (\bar{B^d} - \delta B^d \sigma_i \sigma_j)\Delta_{ij} - \Gamma^d \sum_{i\in G^{\star}}^{K} \sigma_i. \qquad (5)$$

From this point of view, T_d is just a parameter which controls the quality of design: when $T_d \rightarrow \infty$, the ensemble of designed sequences is essentially the same as random, while at lower T_d designed sequences are statistically strongly biased by optimization the design energy.

The protein freezing transition and sequence design have been the subject of a large number of works—for example, [14–17].

We work on artificial model proteins design. *In vivo* De Novo design of globular protein has long been successfully carried out (see, for example, [18] for a four-helix protein design from first principles). The off-lattice model sequence design works [19,20] on a 60-residue SH3 protein domain showed that the folding quality of designed sequences varies for this small protein.

C. High-T_d expansion of the sequence-averaged target state energy and random energy model

First let us look at the energy of target conformation $\star$, averaged over the ensemble of designed sequences:

$$\langle E_\star(\text{seq}) \rangle = \sum_{\text{seq}} P_{\text{seq}}^\star H(\text{seq}, \star). \tag{6}$$

As in [12], we can evaluate this to the lowest order in the high-T_d expansion:

$$\langle E_\star(\text{seq}) \rangle = \langle H \rangle + \frac{1}{T_d} (\langle H^d H \rangle - \langle H^d \rangle \langle H \rangle)$$

$$= N\bar{B} - \frac{1}{T_d} (N \delta B \, \delta B^d + K \Gamma \Gamma^d). \tag{7}$$

We see that sequence design, on average, leads to a lowering of the target-state energy. The only novelty, albeit quite trivial, compared at the volume approximation [12], is that both contact energy and surface energy terms contribute to the target energy decrease.

A priori, one could think that the design lowers not only the target-state energy, but also energies of some other states—particularly those similar to the target state. It is well known, however, that because of the geometry of compact conformations, there are not many sufficiently similar compact conformations and, therefore, the statistics of energies of other conformations, to a good approximation, remains unaffected by the design (see [21] for a further, more detailed discussion of this point). This approximation is equivalent to the REM. We will work within this REM approximation.

Technically, the REM can be justified as a mean-field approximation [7] valid for sufficiently stiff polymers. We therefore will assume that chain entropy per monomer, s (defined such that e^{sN} is the number of conformations), is small: $s \ll 1$. In fact, it is known [7,21] that the REM really works at a softer condition $s \lesssim 1$, and in reality it does meet the latter condition (it is estimated to be about 0.4 in proteins).

III. MONOMER DISTRIBUTIONS INSIDE AND ON THE SURFACE

A. Design by solvation

As stated above, our major goal is to address surface effects in terms of the distribution of σ values among surface monomers. With the designed sequences, we can first ask, what is the distribution of surface monomers $p^\star(\sigma)$ just in the design state when conformation is frozen? This is of course fundamentally a simple question, because the statistical mechanics of sequences at quenched conformation is not frustrated [21]. Basically, what one has to do is to take the probability distribution of sequences (4) and to integrate out all spin variables $\{\sigma_i\}$ except surface monomers—i.e., all i except those belonging to the target conformation surface set $G^\star$.

As a warm-up, let us perform this procedure for the simple yet important special case of a design Hamiltonian, in which we set $\delta B^d = 0$. In this case, design affects only surface monomers. The probability distribution of sequences is tremendously simplified; it can be factorized:

$$P_{\text{seq}}^\star = P_{\text{seq}}^{(0)} \frac{\exp\left[(\Gamma^d/T_d) \sum\limits_{i \in G^\star}^{K} \sigma_i \right]}{\sum\limits_{\text{seq}'} P_{\text{seq}'}^{(0)} \exp\left[(\Gamma^d/T_d) \sum\limits_{i \in G^\star}^{K} \sigma_i' \right]}$$

$$= \prod_{j \notin G^\star} p(\sigma_j) \prod_{i \in G^\star} p^\star(\sigma_i), \tag{8}$$

which means that the ensemble of monomers in the bulk of the globule remains unaffected, distributed as $p(\sigma)$, while every surface monomer, independently of others, is distributed as

$$p^\star(\sigma) = c \, p(\sigma) \exp(\Gamma^d \sigma / T_d), \tag{9}$$

where c is the normalization factor, $c = 1 / \int p(\sigma') \exp(\Gamma^d \sigma' / T_d) d\sigma'$. Thus, for all monomers, including $N - K$ monomers that were in the bulk during the design process and K monomers that were on the surface, the overall distribution of σ reads

$$p_{\text{tot}}(\sigma) = \frac{(N - K)p(\sigma) + K p^\star(\sigma)}{N} = p(\sigma) + \frac{K}{N}[p^\star(\sigma) - p(\sigma)]. \tag{10}$$

This result, Eq. (9), indicates that even this simplified design procedure, with $\delta B^d = 0$, favors hydrophilic monomers on the surface, because for hydrophilic monomers with $\sigma > 0$, we have $p^\star(\sigma) > p(\sigma)$. Compared with the bare monomer distribution $p(\sigma)$, hydrophilic monomers have a larger probability to appear on the surface of target conformation.

The simplified $\delta B^d = 0$ design scheme is reminiscent of the method used in [10,11], in which all surface monomers of target conformation ($G^\star$ in our notation) are made hydrophilic while all monomers inside the globule are hydrophobic. In our more general consideration, it is just more probable but not necessary for the surface exposed monomers to become hydrophilic during the design. The model of the works [9] corresponds to the $T_d \to 0$ limit of our theory.

B. Surface-monomer distribution in the ground state

Let us continue examination of the simplified design scheme, with $\delta B^d = 0$, when only surface energy biases the choice of sequences (design by solvation). Our goal now is to find the surface energy correction terms of the ground-

state energy and, as the major step in this direction, we need to consider the surface-monomer distribution $f(\sigma)$ in the ground state. One should realize that in the ground state, the set of surface exposed monomers may or may not be similar to the set of monomers exposed to the surface during the design; in other words, $f(\sigma)$ might be similar to $p^\star(\sigma)$ or might be quite different from it. Therefore, there are two contributions to the surface energy, one due to the monomers exposed to the surface and the other due to the fact that selection of surface monomers affects the monomer composition left inside the globule. To express this quantitatively, we write for the arbitrary state an equation similar to Eq. (10):

$$p_{\text{tot}}(\sigma) = p_{\text{in}}(\sigma) + \frac{K}{N}[f(\sigma) - p_{\text{in}}(\sigma)], \qquad (11)$$

where $p_{\text{in}}(\sigma)$ is the distribution of monomers left inside. Comparing Eqs. (10) and (11), we find

$$p_{\text{in}}(\sigma) \simeq p(\sigma) + \frac{K}{N}[p^\star(\sigma) - f(\sigma)]. \qquad (12)$$

As everywhere, we neglect here the terms $O((K/N)^2)$. Thus, we directly see already here how the deviation of the state from the target state comes into play.

To compute $f(\sigma)$ for the ground state, we adapt for designed sequences the procedure which was developed in Ref. [9] for random sequence solvation. To make our work self-contained, we briefly outline major steps.

We begin by constructing a separate REM, called the sub-REM, for each possible choice of surface monomers G. Indeed, for each G there are still many conformations available. The number of such conformations is naturally written in the form $M_G = e^{Ns - K\omega_G}$, where s is conformational entropy per monomer in volume approximation and ω_G is entropy loss due to confinement of some monomers on the surface. Although M_G is much smaller than the total number of conformations, $M = e^{sN}$, but the entropy loss caused by fixation of G monomers on the surface is only a surface effect $O(K)$. Following [9], we adopt a bold approximation that $\omega_G = \bar{\omega}$ is independent of G; in this approximation, counting all states shows that $\bar{\omega} = \ln(Ne/K)$.

For each sub-REM, the energies of all M_G states are random in the sense that they depend on random sequence realization, and it is reasonable to assume [9] that these energies are independent Gaussian variables, because each energy, according to formula (1), has of order-N mutually statistically independent bulk contributions and of order-K independent surface contributions. To write down the resulting Gaussian distribution, we should determine the corresponding mean and variance. The mean is found by averaging the bulk terms $(\bar{B} - \delta B \sigma_i \sigma_j)$ over the distribution $p_{\text{in}}(\sigma)$ plus averaging the surface terms $-\Gamma \sigma_i$ over the distribution $f(\sigma)$, and the variance is similarly found by averaging the second moment. This results finally in the following Gaussian distribution of random energy:

$$w_G(E) \propto \exp\left[- \frac{[E - (N\bar{B} - K\Gamma \gamma_G)]^2}{2N\delta B^2 + 2K\delta B^2 \beta_G} \right], \qquad (13)$$

where

$$\gamma_G = \int \sigma f(\sigma) d\sigma,$$

$$\beta_G = 2 \int \sigma^2 [p^\star(\sigma) - f(\sigma)] d\sigma. \qquad (14)$$

Notice that the dependence on the surface-monomer group G is only through the surface-monomer distribution and that the dependence on design is due to the $p^\star(\sigma)$.

Every sub-REM has a certain ground-state energy $E_g(G) = E_g\{f(\sigma)\}$, which is just the lowest of M_G random energies drawn independently from the distribution (13). Energy $E_g\{f(\sigma)\}$ is still a random variable; its probability distribution can be found from the so-called extreme value statistics [22] (see also the Appendix)

$$\mathcal{W}_G(E) = \frac{1}{T_{\text{fr}}} \exp\left[\frac{\delta E_g}{T_{\text{fr}}} - \exp\left(\frac{\delta E_g}{T_{\text{fr}}} \right) \right], \qquad (15)$$

where $T_{\text{fr}} = \delta B / \sqrt{2s}$ and $\delta E_g = E - E_g^{\text{typ}}\{f(\sigma)\}$ is the deviation of the ground-state energy from its most probable (typical) value, which includes both volume and surface contributions:

$$E_g^{\text{typ}}\{f(\sigma)\} = N(\bar{B} - 2sT_{\text{fr}}) + K(\bar{\omega}T_{\text{fr}} - \Gamma \gamma_G + sT_{\text{fr}}\beta_G). \qquad (16)$$

Notice that the only dependence of the probability distribution $\mathcal{W}_G$ on the surface monomers G is hidden in γ_G and β_G inside the most probable energy $E_g^{\text{typ}}\{f(\sigma)\}$ and the dependence on design is also there inside β_G [see formula (14)]. We also mention that $T_{\text{fr}} = \delta B / \sqrt{2s}$ appearing here as a parameter of the ground-state distribution happens to have its physical meaning—it is volume-approximated freezing temperature of the random sequence polymer [9].

The probability to get the ground-state energy anywhere below its typical most probable value (16) is exponentially small. However, we try exponentially many times—namely, we have to choose the lowest among $e^{K\bar{\omega}} = (Ne/K)^K$ sub-REM ground states. Therefore, we have a good chance to find some particular subgroup G with energy noticeably below the typical value (16). Essentially, what we have to do now is to resort a second time to extreme value statistics and find the expectation value of the lowest among the sub-REM ground states. It is convenient to perform this operation in a slightly different, but equivalent form. Namely, we note that the low-energy tail of the ground-state probability distribution, Eq. (15), is exponential and, therefore, it looks effectively like Boltzmann distribution, with T_{fr} playing the role of temperature.

It is useful to note here that treating the tail of the distribution (15) as effective Boltzmann distribution with temperature T_{fr} is reminiscent and essentially equivalent to the consideration given in Ref. [3] and explaining the origin of the phenomenologically discovered quasi-Boltzmann distri-

bution over the ensemble of evolutionary selected proteins.

Returning to our argument, finding the lowest among the E_g^{typ} of the sub-REMs is equivalent to minimizing the effective "free energy," in which the effective entropy is given by the number of ways to choose K monomers with distribution $f(\sigma)$ from N monomers with distribution $p_{\text{tot}}(\sigma)$:

$$s\{f(\sigma)\} = -\int p_{\text{tot}}(\sigma)[\phi \ln \phi + (1-\phi)\ln(1-\phi)]d\sigma,$$

$$(17)$$

where $\phi(\sigma) = Kf(\sigma)/Np_{\text{tot}}(\sigma)$ has the meaning of the fraction of monomers with type σ that are exposed to the surface. Including this effective entropy, we have now the effective "free energy"

$$E_g^{\text{typ}}\{f(\sigma)\} - T_{\text{fr}}Ns\{f(\sigma)\}.$$

$$(18)$$

We minimize this with respect to $f(\sigma)$, subject to normalization condition $\int f(\sigma)d\sigma = 1$, and obtain

$$f(\sigma) = \frac{N}{K}\frac{p_{\text{tot}}(\sigma)}{1 + \Lambda e^{\eta_{\text{fr}}(\sigma)}},$$

$$(19)$$

where $\eta_{\text{fr}}(\sigma) = 2s\sigma^2 - (\Gamma/T_{\text{fr}})\sigma$ and Λ is the Lagrange multiplier which has to be determined from the normalization condition $\int f(\sigma)d\sigma = 1$. Comparing this with Ref. [9], we see that the only role of design in this case is the modification of the monomer distribution: instead of bare distribution $p(\sigma)$, we have now the modified one $p_{\text{tot}}(\sigma)$. Let us see what the consequences of this replacement fare.

Let us concentrate on the regime without the depletion effect, when $\phi(\sigma) \ll 1$ at all values of σ. This means that for any monomer type, only a small fraction of it is solvated to the surface region. Under such an assumption, $f(\sigma)$ can be approximated as

$$f(\sigma) \propto e^{-\eta_{\text{fr}}(\sigma)}\left\{1 + \frac{K}{N}\left[c\exp\left(\frac{\Gamma^d}{T_d}\sigma\right) - 1\right]\right\}p(\sigma), \quad (20)$$

where we dropped for simplicity the σ-independent normalization factor. To gain some insight, let us look at $f(\sigma)$ for a couple of simple examples of bare monomer probability distributions. Since a real distribution involves a large number (20) of monomer species, we examine two limits of two monomer species and of the continuous Gaussian distribution.

1. Example: Bimodal distribution

In the simplest black-and-white model [23] two types of monomers, one hydrophilic and one hydrophobic, appear with same probability:

$$p(\sigma) = \frac{1}{2}[\delta(\sigma + 1) + \delta(\sigma - 1)].$$
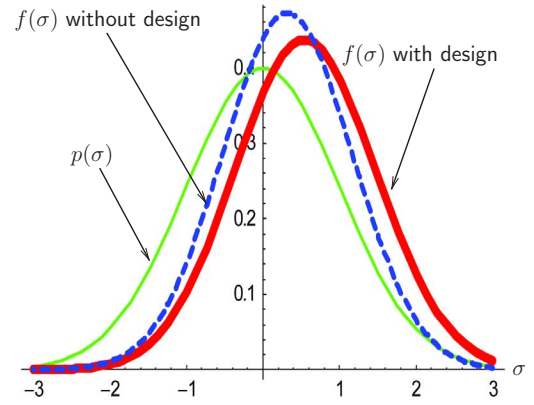
$$(21)$$

Simple calculation shows that

FIG. 2. (Color online) A comparison of the original distribution and surface monomer distribution with design, together with the case of without design. Design favors the monomers with hydrophilic type. For these plots as an example, we used the following values of parameters: $K/N = 0.3$, $s = 0.1$, $\Gamma^d/T_d = 1$, and $\Gamma/\delta B = 1$.

$$f(\sigma) \propto \delta(\sigma + 1)e^{-\Gamma/T_{\text{fr}}}\left[1 - \frac{K}{N}\tanh\left(\frac{\Gamma^d}{T_d}\right)\right] + \delta(\sigma - 1)e^{\Gamma/T_{\text{fr}}}$$

$$\times \left[1 + \frac{K}{N}\tanh\left(\frac{\Gamma^d}{T_d}\right)\right].$$

$$(22)$$

We see that there are two effects bringing hydrophilic monomers to the surface—that is, increasing $f(+1)$ at the expense of decreasing $f(-1)$. The first effect is due to Γ and is measured by the ratio Γ/T_{fr}. This effect is present in random heteropolymers, has nothing to do with design, and is simply energetic: since it is more favorable for the $\sigma > 0$ monomers to be on the surface, so the surface gets enriched with such monomers. The second effect is entirely due to design, and it is governed by the design parameters Γ^d/T_d. This effect is washed away at large design temperature, and it saturates at small T_d. Notice that this design effect is only a surface effect; its maximal possible role is proportional to K/N. This is because the best one can do with this type of design is to shift the monomeric composition by an amount of about K/N.

2. Example: Gaussian distribution

The opposite limit is presented by

$$p(\sigma) = \frac{1}{\sqrt{2\pi}}\exp\left(-\frac{\sigma^2}{2}\right).$$

$$(23)$$

The corresponding surface monomer distribution is the following:

$$f(\sigma) \propto \left(1 + \frac{K}{N}\left\{\exp\left[\frac{\Gamma^d\sigma}{T_d} - \frac{1}{2}\left(\frac{\Gamma^d}{T_d}\right)^2\right] - 1\right\}\right)$$

$$\times \exp\left(-\frac{1+4s}{2}\sigma^2 + \frac{\Gamma}{T_{\text{fr}}}\sigma\right).$$

$$(24)$$

In Fig. 2, we show a plot of an example of $f(\sigma)$ as a function of σ for the Gaussian case. For comparison, random sequence solvation ($T_d \to \infty$) is also included. It can be seen

that design enriches the surface with hydrophilic monomers such that the distribution is shifted toward the hydrophilic region compared to the no-design case, which by itself is already a shift relative to the bare distribution $p(\sigma)$.

C. Sequence design by both solvation and monomer contacts: Mean-field approximation

In the preceding section, we considered sequences designed by the effect of preferential solvation of certain monomer types under the chain preparation conditions. This corresponds to having only the second term in the design Hamiltonian (5) or having $\delta B^d = 0$. By contrast, sequence design by monomer-monomer contacts—i.e., the limit of $\Gamma^d = 0$ and $\delta B^d \neq 0$—was considered in the literature before (see, e.g., review article in [21] and references therein).

In this section, we consider the general case when both volume and surface terms of the design Hamiltonian (5) are present. To make the argument, we resort to the mean-field approximation for the design system. That means that we consider design by an effective field which couples to the variables σ and acts differently on surface and bulk monomers. Since the design Hamiltonian (5) is quadratic in σ, the said "design field" is proportional to $\bar{\sigma}$—an average value of σ defined self-consistently. It is then easy to realize that this field vanishes in the bulk, because in our model design does not affect the overall composition of the chain, and, therefore, $\bar{\sigma}_{\text{bulk}} = 0$. Therefore, the mean-field approximated design Hamiltonian reads

$$H^d_{\text{mean field}} \simeq -(\delta B^d z \bar{\sigma}_{\text{surf}} + \Gamma^d) \sum_{j \in G^\star}^{K} \sigma_j. \quad (25)$$

Here, $\bar{\sigma}_{\text{surf}}$ is the average σ of surface monomers (that is, of monomers which happen to be on the surface during the design process) and z is the coordination number (the number of neighbors) for surface monomers.

Many variants of the mean-field approximation were used in protein physics [3,14–17].

Within the mean-field approximation, the probability distribution of designed sequences (4) gets factorized into independent distributions of all monomers, just like in the $\delta B^d = 0$ case. Accordingly, we obtain a distribution of surface monomers, similar to Eq. (9):

$$p^\star(\sigma) \propto p(\sigma) \exp\left[\frac{\delta B^d z \bar{\sigma}_{\text{surf}} + \Gamma^d}{T_d} \sigma \right], \quad (26)$$

where we dropped for brevity the σ-independent normalization factor. Now, the value of $\bar{\sigma}_{\text{surf}}$ must be determined from the self-consistency condition

$$\bar{\sigma}_{\text{surf}} = \int \sigma p^\star(\sigma) d\sigma = \frac{\int \sigma p(\sigma) \exp\left[\dfrac{\delta B^d z \bar{\sigma}_{\text{surf}} + \Gamma^d}{T_d} \sigma \right] d\sigma}{\int p(\sigma) \exp\left[\dfrac{\delta B^d z \bar{\sigma}_{\text{surf}} + \Gamma^d}{T_d} \sigma \right] d\sigma}. \quad (27)$$

To gain insight into the properties of the latter equation, it is useful to consider examples of bimodal and Gaussian distri-

butions for $p(\sigma)$. We do that a few lines below, in Sec. III D, but here we notice that once $\bar{\sigma}_{\text{surf}}$ is determined, the rest of the analysis follows automatically along the lines of our previous consideration in Sec. III B. Indeed, all we need to know to implement the result (19) is the overall monomer distribution $p_{\text{tot}}(\sigma)$, which is known as soon as $p^\star(\sigma)$ is determined [see Eq. (10)]. Therefore, we can directly use our results, Eqs. (20), (22), and (24), with the replacement $\Gamma^d \rightarrow \Gamma^d + \delta B^d z \bar{\sigma}_{\text{surf}} \equiv \Gamma'^d$.

With that in mind, let us return briefly to the determination of $\bar{\sigma}_{\text{surf}}$.

D. Implementing the self-consistency condition

1. Example: Bimodal distribution

For bimodal $p(\sigma)$, Eq. (27) becomes

$$\bar{\sigma}_{\text{surf}} = \tanh\left[\frac{\delta B^d z \bar{\sigma}_{\text{surf}} + \Gamma^d}{T_d} \right]. \quad (28)$$

At $\Gamma^d = 0$, this equation has nontrivial nonvanishing solutions only if $\delta B^d z / T_d > 1$, in which case there are automatically two solutions of the opposite sign. That means that the nonzero $\bar{\sigma}_{\text{surf}}$ in this case results only from spontaneous symmetry breaking, because without Γ^d the system has no preference for hydrophobic or hydrophilic monomers dominating the surface. The nonzero $\Gamma^d > 0$ breaks this symmetry and yields always one and only one positive solution for $\bar{\sigma}_{\text{surf}}$ (and possibly two negative ones which we ignore because they have higher free energy).

To some extent, the example with bimodal distribution of monomers is just like an Ising ferromagnet; thus, Eq. (28) has a form equivalent to the Ising-model mean-field equation of state [24].

In this bimodal case, we have $\bar{\sigma}_{\text{surf}} < 1$, which means that in the replacement $\Gamma^d \rightarrow \Gamma'^d$, the solvation term Γ^d dominates if $\Gamma^d > \delta B^d z$.

2. Example: Gaussian distribution

For Gaussian $p(\sigma)$, Eq. (27) is easily explicitly resolved:

$$\bar{\sigma}_{\text{surf}} = \frac{\Gamma^d / T_d}{1 - z \delta B^d / T_d}. \quad (29)$$

This is usually very close to Γ^d / T_d, because (see Sec. IV) in the most interesting regime close to the triple point of the phase diagram, the denominator is dominated by unity.

IV. FREE ENERGY AND PHASE DIAGRAM OF DESIGNED POLYMERS

A. Preliminary remarks

In this section, we will consider the possible phases of the heteropolymer whose sequence is designed as discussed above. A similar problem in volume approximation is well known in the literature (see, e.g., the review in [21] and references therein). Specifically, we will consider three phases and the transitions between them. We will summarize the relations between phases in terms of the phase diagram,
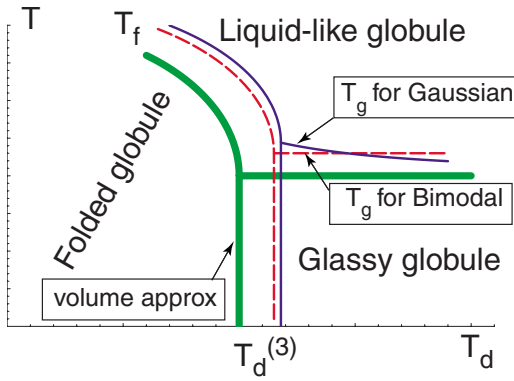
FIG. 3. (Color online) A sketch of the phase diagram of the heteropolymer system. There are three phases: random liquidlike globule, frozen glassy globule, and folded globule. The surface term shifts the phase diagram for volume approximation. For the bimodal distribution, the glass temperature is design independent, while in the Gaussian distribution, the glass temperature increases for better design conditions. Phase diagram lines are not continued to the very-low-T_d region since the system runs out of sequences there.

Fig. 3, in variables T_d and T, which describe, respectively, the ensemble of sequences and the ensemble of conformations for any given sequence. The relevant phases in the diagram are named, respectively, liquidlike globule, glassy globule, and folded globule. We remind the reader that the liquidlike globule is the state where a great many conformations contribute to the partition function; the glassy globule is dominated by one or a few conformations, but those unrelated to the target conformation $\star$; and the folded globule is dominated by the target conformation $\star$. It is fairly obvious and illustrated in Fig. 3 that surface solvation effects do not change the topology of the phase diagram, but do affect the specific positions and shape of the corresponding phase transition lines; these surface-driven changes are the subject of our interest in this section.

The temperatures of the transitions from liquidlike to glasslike and to folded globules are called the glass temperature T_g and folding temperature T_f, respectively. Our goal is to analyze the role of surface solvation effects and design in both T_g and T_f. In other words, we want to calculate how surface corrections to T_f and T_g depend on the design temperature T_d.

As regards the third phase transition line, that between folded and glassy globule phases, this line must be vertical in the phase diagram. Indeed, both folded and glassy globules are zero-entropy states; the transition between them cannot be driven by temperature change. In a phase diagram like in Fig. 3, the corresponding phase boundary must be represented by a line parallel to the temperature axis. This argument does not rely on the volume approximation and, therefore, remains valid independently of surface solvation effects. Therefore, this line of phase transition is entirely described by the value of design temperature at which there is the triple point; we call it $T_d^{(3)}$. We want to calculate also the surface contribution to this quantity.

The way we approach the phase diagram is based on the idea that for any frozen globule phase, whether glassy or folded, the free energy coincides with energy, because entropy vanishes given the number of contributing states of order unity. On the other hand, the free energy of the liquid-like globule we can find due to the property of the REM that the quenched averaged free energy is equal to the annealed average above the glass temperature. Therefore, what we shall do is to compute the annealed average free energy and find the temperature of entropy "catastrophe"—at which entropy vanishes: that is, the glass temperature. Of course, our goal is to address surface terms and design terms in this procedure. Following this program, we write the annealed average as a functional of the surface monomer distribution $f(\sigma)$ as

$$
\begin{aligned}
F\{f\} &= -T \ln \sum_{\text{conf}} e^{-E/T} \\
&\simeq -T \ln \int e^{sN - K\omega_G + Ns\{f(\sigma)\}} w_G(E) e^{-E/T} dE \\
&\simeq \bar{F} + F_{\text{surf}}\{f\},
\end{aligned}
\tag{30}
$$

where

$$
\frac{\bar{F}}{N} = \bar{B} - sT - \frac{\delta B^2}{2T}
\tag{31}
$$

is the bulk contribution to the free energy and

$$
\begin{aligned}
\frac{F_{\text{surf}}\{f\}}{K} &= \bar{\omega}T - \frac{\delta B^2}{T} \int \sigma^2 p^\star(\sigma) d\sigma + \frac{N}{K} T \int d\sigma p_{\text{tot}}(\sigma) \\
&\times \left[\phi\left(\eta(\sigma) - \ln \frac{1-\phi}{\phi} \right) + \ln(1-\phi) \right]
\end{aligned}
\tag{32}
$$

is the surface contribution to the free energy. In Eq. (32),

$$
\eta(\sigma) = \frac{\delta B^2}{T^2} \sigma^2 - \frac{\Gamma}{T} \sigma,
$$

$$
p^\star(\sigma) = \frac{p(\sigma) \exp\left(\dfrac{\Gamma' d}{T_d} \sigma \right)}{\int p(\sigma) \exp\left(\dfrac{\Gamma' d}{T_d} \sigma \right) d\sigma}.
\tag{33}
$$

In deriving the free energy, we have used the saddle point approximation since above glass temperature, the free energy is dominated by the saddle point of the partition function.

Optimizing $F_{\text{surf}}\{f\}$ yields

$$
\phi^\star = \frac{1}{1 + \Lambda e^{\eta(\sigma)}} = \frac{K f^\star(\sigma)}{N p_{\text{tot}}(\sigma)},
\tag{34}
$$

and then the optimal value of $F_{\text{surf}}\{f\}$ is evaluated as

$$\frac{F_{\text{surf}}\{f^\star\}}{KT} = 1 + \frac{\displaystyle\int p_{\text{tot}}(\sigma)\ln\frac{\Lambda e^{\eta(\sigma)}}{1+\Lambda e^{\eta(\sigma)}}d\sigma}{\displaystyle\int \frac{p_{\text{tot}}(\sigma)}{1+\Lambda e^{\eta(\sigma)}}d\sigma}$$

$$-\frac{\delta B^2}{T^2}\int \sigma^2 p^\star(\sigma)d\sigma - \ln\int \frac{\Lambda p_{\text{tot}}(\sigma)}{1+\Lambda e^{\eta(\sigma)}}d\sigma. \tag{35}$$

These results are similar to those obtained in Ref. [9], except we have now nonrandom sequences, as manifested by the dependence on T_d. In that work [9], the solvation effect in random sequences was treated as the response of the globule to the surface solvation "field." The linear response regime is characterized by the statistical independence of disordered parts of surface and volume energies. The saturation regime is characterized by the depletion effect, when preferential solvation of a certain monomer species exhausts these monomers from the globule. For completeness, there is also a narrow range of the so-called weak-response regime. Neither weak-response nor saturation regimes are present for the black-and-white polymer model, with bimodal distribution of monomer types.

B. No depletion: Glass temperature

Let us first consider the case when solvation does not cause the depletion of any monomer type—that is, $\phi(\sigma) \ll 1$ at every σ; then,

$$\frac{F_{\text{surf}}\{f^\star\}}{KT} \simeq -\frac{\delta B^2}{T^2}\int \sigma^2 p^\star(\sigma)d\sigma - \ln\int p_{\text{tot}}(\sigma)e^{-\eta(\sigma)}d\sigma. \tag{36}$$

First we will discuss how the glass temperature T_g is affected by the surface energy, relative to the volume approximated value $T_{\text{fr}}=\delta B/\sqrt{2s}$. The glass temperature must be determined from by the condition $-\frac{\partial F\{f\}}{\partial T}\big|_{T_g}=0$. Since $-\frac{\partial \bar{F}}{\partial T}\big|_{T_{\text{fr}}}=0$, we can write (denoting temperature derivatives by prime sign) $\Delta T_g \simeq T_g - T_{\text{fr}} \simeq -\frac{F'_{\text{surf}}}{\bar{F}''}\big|_{T_{\text{fr}}}$. Then

$$\frac{\Delta T_g}{T_{\text{fr}}} \simeq \frac{K}{N}\left[\frac{\Gamma}{2sT_{\text{fr}}}\frac{\displaystyle\int \sigma p(\sigma)e^{-\eta_{\text{fr}}(\sigma)}d\sigma}{\displaystyle\int p(\sigma)e^{-\eta_{\text{fr}}(\sigma)}d\sigma} + \int \sigma^2 p^\star(\sigma)d\sigma\right.$$

$$\left. - \frac{2\displaystyle\int \sigma^2 p(\sigma)e^{-\eta_{\text{fr}}(\sigma)}d\sigma}{\displaystyle\int p(\sigma)e^{-\eta_{\text{fr}}(\sigma)}d\sigma} - \frac{1}{2s}\ln\int p(\sigma)e^{-\eta_{\text{fr}}(\sigma)}d\sigma\right], \tag{37}$$

where we have replaced $p_{\text{tot}}(\sigma)$ with $p(\sigma)$ because ΔT_g itself is already of order $O(K/N)$ and we neglect any higher-order corrections. The design effect is present here through $p^\star(\sigma)$.

Relative to the no-solvation case, the order-$O(K/N)$ correction to the glass temperature is positive; in other words,

the glass temperature is increased due to the solvation effect, and thus the designed heteropolymers may be found in the glassy globule state in a broader temperature range.

In the following part, we will further simplify and discuss ΔT_g using the examples of $p(\sigma)$.

1. Example: Bimodal distribution

For bimodal $p(\sigma)$, $\int \sigma^2 p^\star(\sigma)d\sigma=1$, and design effect does not show up in the glass temperature,

$$\Delta T_g = \frac{K}{2sN}T_{\text{fr}}\left[\frac{\Gamma}{T_{\text{fr}}}\tanh\left(\frac{\Gamma}{T_{\text{fr}}}\right) - \ln\cosh\left(\frac{\Gamma}{T_{\text{fr}}}\right)\right]$$

$$\simeq \frac{K}{N}\begin{cases}\dfrac{\Gamma^2}{2\sqrt{2s}\,\delta B} & \text{when } \dfrac{\Gamma}{T_{\text{fr}}}\ll 1, \\[3mm] \dfrac{\ln 2}{(2s)^{3/2}}\delta B & \text{when } \dfrac{\Gamma}{T_{\text{fr}}}\gg 1. \end{cases} \tag{38}$$

When the solvation strength Γ is small, $\Gamma/T_{\text{fr}}\ll 1$, this corresponds to the statistically independent region, in which the surface term and volume term in a heteropolymer globule are roughly statistically independent. In this region, $\Delta T_g \propto \Gamma^2$.

The region of $\Gamma/T_{\text{fr}}\gg 1$ is the saturation region. When the solvation strength Γ is so large, essentially all the surface monomers are of hydrophilic type, as can be clearly seen from the surface-monomer distribution, Eq. (22): when $\Gamma/T_{\text{fr}}\gg 1$, the surface-monomer distribution function is dominated by the hydrophilic term $f(+1)$. In this regime, ΔT_g becomes independent of Γ, because Γ is already so large that all surface places are occupied by hydrophilic monomers and further increase of Γ cannot change anything.

2. Example: Gaussian distribution

For Gaussian $p(\sigma)$, we can write $\int \sigma^2 p^\star(\sigma)d\sigma = (\Gamma'^d/T_d)^2+1 \simeq (\Gamma^d/T_d)^2(1+2z\delta B^d/T_d)+1 \simeq (\Gamma^d/T_d)^2+1$, where the asymptotic form comes from the fact that $T_d/(z\delta B^d)\gg 1$ since we work in the regime of $T_d > T_d^{(3,0)} = \delta B^d/\sqrt{2s}$, where $T_d^{(3,0)}$ is the triple point in volume approximation. Then we have

$$\Delta T_g \simeq \frac{K}{N}T_{\text{fr}}\left[\left(\frac{\Gamma^d}{T_d}\right)^2 + 6s + \frac{\Gamma^2}{2\delta B^2}\right], \tag{39}$$

where we also used the fact that $s\ll 1$. As in Ref. [9], the system with Gaussian-distributed σ, unlike the bimodal one, has the weak-response regime at very small Γ, when $\Gamma/T_{\text{fr}} \ll s$; in this regime, surface solvation is insignificant. The region of $\Gamma/T_{\text{fr}} \gg \sqrt{24}s$ is the regime where volume and surface disorder are statistically independent, and the result in this region, in terms of dependence on Γ, is similar to that of the bimodal distribution.

Of course, the major difference from the bimodal example is that in Gaussian case, there is an additional term due to sequence design. That term increases the glass temperature, which means design, as usual, makes for a more stable ground state.

C. Triple point

Now let us consider the folded region and begin with the surface-corrected triple point $T_d^{(3)}$; we want to see its change $\Delta T_d^{(3)} = T_d^{(3)} - T_d^{(3,0)}$ relative to $T_d^{(3,0)} = \delta B^d / \sqrt{2s}$ determined in volume approximation. In general, a triple point is determined from the condition that the glass temperature equal the folding temperature, $T_f = T_g$. The glass temperature, along with its surface corrections, is already known to us; see formula (37) or its simplified versions (38) and (39). The folding temperature T_f should be calculated from $\langle E_\star(\text{seq}) \rangle = F\{f^\star\}$. We take the averaged ground-state energy from formula (7), and we take the $F\{f^\star\}$ from Eqs. (31) and (35); in the latter (which is the surface part), we must neglect all order-$O(K/N)$ corrections. The result reads

$$\Delta T_d^{(3)} \simeq \frac{K}{N} T_d^{(3,0)} \left[\frac{\Gamma \Gamma^d}{\delta B \, \delta B^d} - \int \sigma^2 p^\star(\sigma) d\sigma \right.$$
$$\left. - \frac{1}{2s} \ln \int p(\sigma) e^{-\eta_{\text{fr}}(\sigma)} d\sigma \right]. \quad (40)$$

From this formula, it is not even clear whether $\Delta T_d^{(3)}$ is positive or negative. As with other cumbersome results, let us look at the specific examples of $p(\sigma)$.

1. Example: Bimodal distribution

For the bimodal distribution $p(\sigma)$, we have

$$\Delta T_d^{(3)} \simeq \frac{K}{N} T_d^{(3,0)} \left[\frac{\Gamma \Gamma^d}{\delta B \, \delta B^d} - \frac{1}{2s} \ln \cosh\left(\frac{\Gamma}{T_{\text{fr}}} \right) \right]$$
$$\simeq \begin{cases} \dfrac{K}{N} T_d^{(3,0)} \dfrac{\Gamma}{\delta B} \left(\dfrac{\Gamma^d}{\delta B^d} - \dfrac{\Gamma}{2 \delta B} \right) & \text{when } \dfrac{\Gamma}{T_{\text{fr}}} \ll 1, \\[3mm] \dfrac{K}{N} T_d^{(3,0)} \dfrac{\Gamma}{\delta B} \left(\dfrac{\Gamma^d}{\delta B^d} - \dfrac{1}{\sqrt{2s}} \right) & \text{when } \dfrac{\Gamma}{T_{\text{fr}}} \gg 1. \end{cases}$$
$$(41)$$

We see that the design effect increases $\Delta T_d^{(3)}$ and pushes the triple point to the right in the phase diagram Fig. 3, while the solvation effect acts in the opposite direction. Interestingly, in the statistically independent regime, when $\Gamma / T_{\text{fr}} \ll 1$, the sign of $\Delta T_d^{(3)}$ is determined by the competition of the design term $\Gamma^d / \delta B^d$ and folding term $\Gamma / \delta B$. Specifically, large design solvation strength $\Gamma^d / \delta B^d \gg \Gamma / \delta B$ would make $\Delta T_d^{(3)} > 0$ and in this sense the design makes the folded state more stable. From Eq. (38), we already know that the glass temperature is independent of Γ in the saturation region $\Gamma / T_{\text{fr}} \gg 1$. Not surprisingly, in this region, the sign of $\Delta T_d^{(3)}$ is also independent of Γ.

2. Example: Gaussian distribution

When $p(\sigma)$ is Gaussian, we get

$$\Delta T_d^{(3)} \simeq \frac{K}{N} T_d^{(3,0)} \left[\frac{\Gamma \Gamma^d}{\delta B \, \delta B^d} - \frac{\Gamma^2}{2 \delta B^2} \right.$$
$$\left. - 2s - 2s \left(\frac{\Gamma^d}{\delta B^d} \right)^2 (1 + 2z\sqrt{2s}) \right]. \quad (42)$$

The interesting and rather unexpected result is that the design effect, when it is *very* strong, might lead to a reduction of $T_d^{(3)}$; in other words, it might have an adverse effect on the stability of the folded phase. Inspection of the origin of the negative term $\propto -(\Gamma^d / \delta B^d)^2$ shows that its origin is due to the fact that a very strong solvation effect in the design brings in a significant fraction of very strongly solvophilic monomers; even though only a small fraction of them subsequently turn out inside the globule in the folded state, they nevertheless make the destabilizing effect on the globule. We emphasize that such danger exists only when the solvation effect in the design is so strong that not only $\Gamma^d / \delta B^d > \Gamma / \delta B$, but $s\Gamma^d / \delta B^d > \Gamma / \delta B$. It is unclear if such a situation is realistic.

D. Folding temperature

Next let us consider the folding temperature T_f away but not far from the triple point. When $T_d < T_d^{(3)}$, we have

$$\frac{\delta B \, \delta B^d + \dfrac{K}{N} \Gamma \Gamma^d}{T_d} = \left. \left(sT + \frac{\delta B^2}{2T} + F_{\text{surf}} \right) \right|_{T = T_f}. \quad (43)$$

In the vicinity of the triple point, when $T_d = T_d^{(3)} - \Delta T_d$, where $\Delta T_d \ll \frac{K}{N} T_d^{(3)}$, we have $F_{\text{surf}}|_{T = T_f} \simeq F_{\text{surf}}|_{T = T_g, T_d = T_d^{(3,0)}}$ or

$$\frac{\delta B \, \delta B^d}{T_d^{(3)}} \frac{\Delta T_d}{T_d^{(3)}} \simeq sT_f + \frac{\delta B^2}{2T_f} - sT_g - \frac{\delta B^2}{2T_g}, \quad (44)$$

which yields upon some algebra

$$T_f \simeq T_g \left[1 + \sqrt{\frac{\delta B \, \delta B^d}{sT_d^{(3)} T_g} \left(1 - \frac{T_d}{T_d^{(3)}} \right)} \right]. \quad (45)$$

In Fig. 3, the phase diagram of the system is sketched. In the regime of temperature T below T_f, the designed sequences will be thermodynamically stable when folded to the target state. In the regime of $T_d > T_d^{(3)}$, the sequences obtained will be either in a frozen glassy state or in a random liquidlike globule.

E. Depletion effect

In the preceding sections, we assumed no depletion effect. Here we will see what happens if there is depletion. Depletion of monomers may happen for the Gaussian $p(\sigma)$, while in the bimodal case, there is no depletion since the number of monomers for each monomer type is abundant. When depletion occurs, $\phi = 1$ for $\sigma \geq \sigma_m$ and $\phi = 0$ for $\sigma < \sigma_m$, and this leads to

$$\int_{\sigma_m}^{\infty} p_{\text{tot}}(\sigma) d\sigma = K/N. \quad (46)$$

The integration result is

$$\underbrace{\mathrm{erfc}\left(\frac{\sigma_m - \dfrac{\Gamma'^d}{T_d}}{\sqrt{2}}\right) - \mathrm{erfc}\left(\frac{\sigma_m}{\sqrt{2}}\right) - 2}_{>0} = -\frac{N}{K}\,\mathrm{erfc}\left(\frac{\sigma_m}{\sqrt{2}}\right). \tag{47}$$

For the case of no design, $\Gamma'^d = 0$,

$$\mathrm{erfc}\left(\frac{\sigma_m^0}{\sqrt{2}}\right) = \frac{2K}{N}. \tag{48}$$

Therefore, we have $\mathrm{erfc}(\sigma_m/\sqrt{2}) < \mathrm{erfc}(\sigma_m^0/\sqrt{2})$, and this gives $\sigma_m > \sigma_m^0$, so with design, the surface monomers are more hydrophilic than without design. This makes physical sense, and this is also consistent with the no-depletion case $\phi \ll 1$, in which design favors the hydrophilic monomers on the surface.

V. SEQUENCE-SPACE ENTROPY

Sequence design, when it is realized computationally or if it could be realized experimentally, helps finding sequences with particularly low ground-state energy. But of course there is a limit—there is always a sequence whose ground-state energy is the lowest among all sequences, and, therefore, no design can possibly produce any sequence with lower energy. More generally and more practically, the lower the ground-state energy we want to obtain, the fewer sequences exist which can meet our demand. One may want to know how many sequences are there to choose from with any given ground-state energy. The design paradigm provides the general method to solve a such problem. Indeed, we can compute the sequence-space entropy (which is just the logarithm of the number of relevant sequences) as a function of T_d, and as we also know the average ground-state energy as a function of T_d, we can determine the number of sequences as depends on their ground-state energy. This procedure in volume approximation is described in Ref. [21]. Here we want to consider how it is affected by the surface solvation effect.

In principle, the sequence-space entropy depends quite strongly on the target-state fold here denoted as $\star$; this dependence is called the designability of the fold (see, for example, recent work on this subject [25]). Here, we will neglect this fact. This is not because designability is unimportant—it is very important indeed—but our goal is to look at the role of the sequence solvation effect, and so to make this task manageable, we have to sacrifice the designability issue as a zeroth approximation.

To find the sequence-space entropy, we consider the sequence-space free energy $-T_d \ln Z$, where $Z = \sum_{\mathrm{seq}} \exp[-H^d(\mathrm{seq}, \star)/T_d]$. Note that design is to a certain conformation $\star$, so in the partition function here, the summation runs over sequences. The sequence-space entropy per monomer, s_{seq}, can then be found using high-T_d expansion just in the same way as the calculation of $\langle E_\star(\mathrm{seq})\rangle$, formula (7). The result is given by

$$s_{\mathrm{seq}} = -\frac{\partial(-T_d \ln Z)}{\partial T_d} \simeq \ln q - \frac{\delta B^{d2} + \dfrac{K}{N}\Gamma^{d2}}{2T_d^2}, \tag{49}$$

where q is the effective number of "letters in the alphabet" determined from the total number of possible sequences of length N: $\mathcal{N}_{\mathrm{seq}} = q^N$. It is not difficult to show that $q = -\sum_\sigma p(\sigma) \ln p(\sigma)$ [see Eq. (3); $q \approx 18$ for real proteins].

The number of sequences is maximal if we impose no constraints on the quality of design, which means the sequence entropy has to be maximal when T_d is at the triple point. Therefore, we can compute $s_{\mathrm{seq}}^{\mathrm{max}}$ using formula (49) at $T_d = T_d^{(3)}$. The result reads

$$s_{\mathrm{seq}}^{\mathrm{max}} \simeq \ln q - s\left[1 - \frac{2\Delta T_d^{(3)}}{T_d^{(3,0)}} + \frac{K\Gamma^{d2}}{N\delta B^{d2}}\right], \tag{50}$$

where the ratio $\Delta T_d^{(3)}/T_d^{(3,0)}$ should be taken from Eq. (40) or from the simplified versions of it, Eq. (41) or (42).

First, in the volume approximation, when there is no surface term, we have $s_{\mathrm{seq}}^{\mathrm{max}} = \ln q - s$. This result is a very natural consequence of our neglect of the difference in designabilities between different folds. Indeed, the volume approximation of $s_{\mathrm{seq}}^{\mathrm{max}}$ indicates that the number of sequences that can be designed for a given conformation $\star$ is $e^{Ns_{\mathrm{seq}}^{\mathrm{max}}} = \mathcal{N}_{\mathrm{seq}}/\mathcal{N}_{\mathrm{conf}}$, which means that all $\mathcal{N}_{\mathrm{seq}} = q^N$ sequences are equally distributed between $\mathcal{N}_{\mathrm{conf}} = e^{sN}$ conformations. This is because the fraction of sequences with ground-state energy above $\langle E_\star(\mathrm{seq})\rangle$, Eq. (7), is extremely small (see the Appendix), so practically all sequences have their ground-state energy around $\langle E_\star(\mathrm{seq})\rangle$.

Second, we look at the role of the surface effect. For simplicity, we restrict consideration to the most typical regime of statistical independence between surface and bulk contributions. For both the bimodal distribution and Gaussian distribution, plugging $\Delta T_d^{(3)}$ into Eq. (50), we get the following simple result:

$$s_{\mathrm{seq}}^{\mathrm{max}} = \ln q - s - \frac{sK}{N}\left[\frac{\Gamma}{\delta B} - \frac{\Gamma^d}{\delta B^d}\right]^2. \tag{51}$$

This tells us that the surface solvation effect reduces the number of sequences, $s_{\mathrm{seq}}^{\mathrm{max}} < \ln q - s$. This happens because some of the sequences, with inadequate supply of hydrophilic monomers, have their ground-state energies above $\langle E_\star(\mathrm{seq})\rangle$ when we look at them carefully enough to notice their surface energy. Accordingly, the fraction of sequences with ground-state at $\star$ is below its "fair" share of $e^{(\ln q - s)N}$ and even the maximal sequence-space entropy falls short of its volume-approximated value $\ln q - s$. Only very careful design, at which $\Gamma^d/\delta B^d = \Gamma/\delta B$, would be able to provide the ensemble of sequences adequate to their solvation conditions, in which case the solvation effect does not increase energy and does not rule any sequences out of competition. Notice that the condition $\Gamma^d/\delta B^d = \Gamma/\delta B$ does not involve the design temperature; it specifies only the balance of solvation and bulk heteropolymeric effects.

Let us now look at the situation differently; namely, let us write down the folding temperature T_f in terms of s_{seq} instead

of T_d. Indeed, T_d is a purely technical concept which may or may not directly correspond to experimental reality; for instance, design can be controlled by some analog of the solvent quality instead of temperature. At the same time, the sequence entropy is a very clear quantity; it is the number of sequences whose ground-state stability corresponds to the temperature T_f. Simple algebra shows that

$$T_f = T_g \left[1 + \sqrt{\frac{\delta B\, \delta B^d}{s T_g T_d^{(3)}} \left(1 - \sqrt{\frac{\ln q - s_{\text{seq}}^{\max}}{\ln q - s_{\text{seq}}}} \right)} \right]. \quad (52)$$

This allows us to reinterpret the phase diagram, Fig. 3, with sequence entropy on the horizontal axis.

Finally, it is known [8] that the quality of design is best characterized by the energy gap between the energy of the sequence in its purported target state and the average ground-state energy:

$$\Delta \epsilon = \frac{F\{f^\star\}|_{T_g} - \langle E_\star(\text{seq}) \rangle}{N}. \quad (53)$$

Therefore, we should look at the relation between sequence entropy and $\Delta \epsilon$. From the above results, we have found

$$s_{\text{seq}} = \ln q - s \left(1 + \frac{\Delta \epsilon}{\sqrt{2s}\, \delta B} + \frac{K}{N} \zeta \right)^2 \left(1 + \frac{K}{N} \xi \right), \quad (54)$$

where the solvation-related coefficients are given by

$$\xi = \frac{\Gamma^{d^2}}{\delta B^{d^2}} - 2 \frac{\Gamma \Gamma^d}{\delta B\, \delta B^d} \quad (55)$$

and

$$\zeta = \int \sigma^2 p^\star(\sigma) d\sigma + \frac{1}{2s} \ln \int p(\sigma) e^{-\eta_{\text{fr}}(\sigma)} d\sigma. \quad (56)$$

There are fewer sequences available for larger energy gap design.

1. Example: Bimodal distribution
When $p(\sigma)$ is bimodal,

$$\zeta = \frac{1}{2s} \ln \cosh \frac{\Gamma}{T_{\text{fr}}}. \quad (57)$$

2. Example: Gaussian distribution
When $p(\sigma)$ is Gaussian,

$$\zeta = \frac{\Gamma^2}{2\delta B^2} + 2s + 2s \left(\frac{\Gamma^d}{\delta B^d} \right)^2. \quad (58)$$

In either case, we see that the number of available sequences drops dramatically as we increase their desired quality by choosing a larger $\Delta \epsilon$.

VI. CONCLUSION

In this paper, we examined the interplay of surface solvation effects and sequence design for a proteinlike heteropolymer globule. Ideologically, our treatment of disordered sequences followed the theoretical studies of heteropolymer folding in Refs. [4–6,12], and in our treatment of preferential solvation we used the approach of Ref. [9]. What we did is we applied the REM in the new challenging context.

As in the volume approximation, designed sequences in the target conformation have lower energy than random sequences. This is not surprising: this is after all the sole purpose of design. Less obvious, we found that the role of preferential solvation for the design itself might be controversial. The problem is that when design conditions favor too strongly the hydrophilicity of the surface monomers, these monomers can have an adverse effect on the overall composition of the sequence and then some hydrophilic monomers are found in the interior, thus making the globule less stable.

Speaking about the phase diagram of the heteropolymer globule, we found that the surface solvation effect operates differently for the two most typical examples of monomer composition. If there are only two types of monomers, then the glass transition temperature remains independent of the design condition, as was found in the volume approximation. But this is no longer the case when there is a wide Gaussian distribution of monomer types; in this case, design brings in a noticeable fraction of very hydrophilic monomers from the tail of the hydrophilicity distribution, and they do affect the glass transition.

To conclude, our study shows it possible to incorporate preferential solvation effects into REM-based heteropolymer theory, and some of the obtained results are quite delicate and unexpected. In reality, the role of surface in molecules of realistic sizes is quite significant, so the effects which were examined here on a perturbative level, considering surface contributions $O(K/N)$ very small, might be quite substantial and very important.

APPENDIX: PROBABILITY DISTRIBUTIONS OF THE LOW-ENERGY STATES IN THE REM

To make this work self-contained, we review here the probability distributions of the low-lying states in the REM. Further details on this subject can be found in [22].

Ground-state energy

Consider some $\mathcal{M} (\mathcal{M} = e^{sN})$ statistically independent energy levels, each distributed with probability density $w(E)$. The question is this: what is the probability distribution $\mathcal{W}(E)$ of the lowest among these $\mathcal{M}$ levels? The general answer, due to the statistical independence, reads

$$\mathcal{W}(E) = \mathcal{M} w(E) \left(\int_E^\infty w(E') dE' \right)^{\mathcal{M}-1}. \quad (A1)$$

Here, $w(E)$ is the probability that there is a level at E, $(\int_E^\infty w(E') dE')^{\mathcal{M}-1}$ is the probability that all other $\mathcal{M} - 1$ levels happen to be above E, and the factor $\mathcal{M}$ reflects the fact that any of the $\mathcal{M}$ levels may play the role of the lowest one.

With a large number $\mathcal{M} \gg 1$ of levels taken from the distribution $w(E)$, the lowest one of them will surely be located somewhere in the low-energy tail of the distribution $w(E)$. In this region, $\int_E^\infty w(E')dE'$ is very close to unity or, in other words, $\int_E^\infty w(E')dE' = 1 - \int_{-\infty}^E w(E')dE' \simeq \exp[-\int_{-\infty}^E w(E')dE']$. Neglecting also 1 compared at $\mathcal{M}$ in $\mathcal{M}-1$, we then rewrite Eq. (A1):

$$\mathcal{W}(E) \simeq \mathcal{M}w(E) \exp\left(-\mathcal{M}\int_{-\infty}^E w(E')dE'\right). \quad (A2)$$

Where is the maximum of $\mathcal{W}(E)$? In other words, what is the most probable ground-state energy E_m? The condition $\mathcal{W}'(E)=0$ yields

$$\mathcal{M}w(E_m) = [\ln w(E_m)]'. \quad (A3)$$

Apart from the logarithmic corrections, this returns the familiar condition traditionally written in a careless form $\mathcal{M}w(E) \simeq 1$ in which units do not match.

We now remember that $w(E)$ is Gaussian,

$$w(E) = \frac{1}{\sqrt{2\pi N \delta B^2}} e^{-E^2/2N\delta B^2}. \quad (A4)$$

In this case formula (A3) reads

$$\mathcal{M}w(E_m) = -E_m/N\delta B^2, \quad (A5)$$

which then implies

$$E_m = -\sqrt{2N\delta B^2 \left\{ \underbrace{\ln \mathcal{M}}_{Ns} - \ln\left[\frac{-E_m}{N\delta B^2}\sqrt{2\pi N\delta B^2}\right] \right\}} \quad (A6)$$

and by simple iteration we finally have

$$E_m \simeq -\sqrt{2s}N\delta B + \frac{\delta B}{2\sqrt{2s}}\ln(4\pi Ns) \simeq -\sqrt{2s}N\delta B + O(\ln N). \quad (A7)$$

The leading term here corresponds to dropping the preexponential factor of $w(E)$ when writing $\mathcal{M}w(E) \simeq 1$, in which case, of course, units match and the result is correct.

It is fairly obvious, and will be verified instantly, that $\mathcal{W}(E)$ is concentrated around E_m. Accordingly, let us assume $E = E_m + \epsilon$, with $|\epsilon| \ll |E_m|$. In this small-ϵ range, we can simplify the Gaussian $w(E)$:

$$\mathcal{M}w(E) = \frac{\mathcal{M}}{\sqrt{2\pi N\delta B^2}} e^{-(E_m+\epsilon)^2/2N\delta B^2}$$

$$\simeq \underbrace{\mathcal{M}w(E_m)}_{\sqrt{2s}/\delta B} e^{\epsilon E_m/N\delta B^2}$$

$$\simeq \frac{1}{T_{\mathrm{fr}}} e^{\epsilon/T_{\mathrm{fr}}}, \quad (A8)$$

where $T_{\mathrm{fr}} = \frac{\delta B}{\sqrt{2s}}$. We further use the Gaussian shape of $w(E)$ and the proper asymptotics of the error function to write

$$\mathcal{M}\int_{-\infty}^E w(E')dE' \simeq \mathcal{M}w(E)\frac{N\delta B^2}{-E} \simeq e^{\epsilon/T_{\mathrm{fr}}}. \quad (A9)$$

Plugging both Eqs. (A8) and (A9) into Eq. (A2), we arrive at the following probability distribution of the ground-state energy:

$$\mathcal{W}(E) \simeq (1/T_{\mathrm{fr}})e^{(\epsilon/T_{\mathrm{fr}})-\exp(\epsilon/T_{\mathrm{fr}})}, \quad (A10)$$

where, once again,

$$\epsilon = (E - E_m), \quad T_{\mathrm{fr}} = \frac{\delta B}{\sqrt{2s}}. \quad (A11)$$

The probability distribution $\mathcal{W}(E)$ is essentially the Gumbel distribution in extreme value statistics [22]. As expected, $\mathcal{W}(E)$ is not symmetric; it decays much faster to the right (higher energies) than to the left (lower energies). However, the characteristic scale in both cases is set by the quantity $T_{\mathrm{fr}} = \delta B/\sqrt{2s}$ and does not contain N in any form. It is only the parameter-free "shape" $e^{\xi-e^\xi}$ that is asymmetric. It is also satisfying that $\mathcal{W}(E)$, Eq. (A10), is correctly normalized.

Lowest- and second-lowest-energy states: Joint distribution

Consider now the joint probability distribution of the two lowest energies. Exact formula for this joint distribution, similar to Eq. (A1), reads

$$\mathcal{W}(E_1, E_2) = \mathcal{M}(\mathcal{M}-1)w(E_1)w(E_2)\left(\int_{E_2}^\infty w(E')dE'\right)^{\mathcal{M}-2}. \quad (A12)$$

Similar to Eq. (A1), we rely here on the statistical independence of the energy levels in the REM, which implies that we should consider the independent factors of one energy level present at E_1 [which gives $w(E_1)$], another energy levels present at E_2 [yielding $w(E_2)$], times the probability that all other $\mathcal{M}-2$ energy levels are above E_2, and times the combinatorial factor $\mathcal{M}(\mathcal{M}-1)$, which reflects the idea that any two of the $\mathcal{M}$ levels can play the role of the lowest and second lowest.

The first step to simplify this distribution is to notice that both the lowest-E_1 and second-lowest-E_2 energy levels are practically always located far in the lower-energy tail of the distribution $w(E)$. Similar to Eq. (A2), we then write

$$\mathcal{W}(E_1, E_2) = \mathcal{M}^2 w(E_1)w(E_2)\exp\left[-\mathcal{M}\int_{-\infty}^{E_2} w(E')dE'\right]. \quad (A13)$$

The next step is to rely on Eqs. (A8) and (A9). Denoting

$$E_1 = E_m + \epsilon_1, \quad E_2 = E_m + \epsilon_2, \quad (A14)$$

and assuming $|\epsilon_{1,2}| \ll |E_m|$, we arrive at

$$\mathcal{W}(E_1, E_2) \simeq \frac{1}{T_{\mathrm{fr}}^2} e^{(\epsilon_1/T_{\mathrm{fr}})+(\epsilon_2/T_{\mathrm{fr}})-\exp(\epsilon_2/T_{\mathrm{fr}})}. \quad (A15)$$
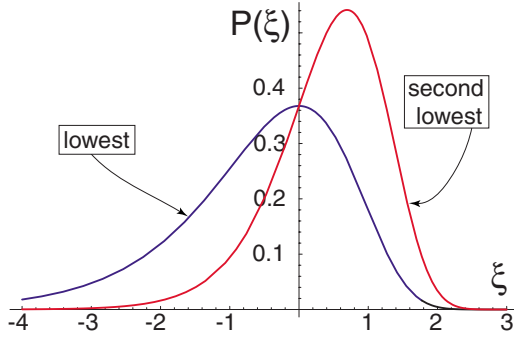
FIG. 4. (Color online) Probability distributions for the lowest-
and second-lowest-energy states in the REM. The horizontal axis is
$\xi = \epsilon/T_{\mathrm{fr}}$, and the vertical axes is the probability density for this
quantity ξ (so the area under each curve in the figure is unity).

Some derivative probability distributions

It is a wonderfully rewarding exercise to establish that
integrating out the first excited state E_2 returns the ground-
state distribution (A10), as it must: $\int_{E_1}^{\infty} \mathcal{W}(E_1,E_2)dE_2$
$= \mathcal{W}(E_1)$.

Probability distribution of the second lowest state

For the second lowest energy, we have by simple integra-
tion

$$W_2(E) = \int_{-\infty}^{E} \mathcal{W}(E_1,E)dE_1 = \frac{1}{T_{\mathrm{fr}}} e^{2(\epsilon/T_{\mathrm{fr}})-\exp(\epsilon/T_{\mathrm{fr}})}.$$

$$(A16)$$

This distribution is different from that of the lowest-energy
state, Eq. (A10), as is clear from Fig. 4. It is hardly a surprise
that the second-lowest-energy distribution is somewhat
shifted towards larger energies compared with the distribu-
tion of the ground state.

Conditional probability of E_2 at given E_1

Suppose ground-state energy is fixed at E_1. What is the
probability distribution of E_2 at the given E_1? According to
the general rules of probabilities, this quantity is equal to

$$\mathcal{W}(E_2|E_1) = \frac{\mathcal{W}(E_1,E_2)}{\mathcal{W}(E_1)} = \frac{1}{T_{\mathrm{fr}}} e^{(\epsilon_2/T_{\mathrm{fr}})-\exp(\epsilon_2/T_{\mathrm{fr}})+\exp(\epsilon_1/T_{\mathrm{fr}})}.$$

$$(A17)$$

For understanding, it is useful to mention that this quantity is
normalized by the condition $\int_{E_1}^{\infty} \mathcal{W}(E_2|E_1)dE_2=1$.

Probability distribution for the gap between lowest and second lowest states

As regards the gap between two lowest-energy states,
$\Delta E = E_2 - E_1$, its probability distribution is also easily found:

$$\mathcal{W}(\Delta E) = \int_{-\infty}^{\infty} dE_1 \int_{E_1}^{\infty} dE_2 \mathcal{W}(E_1,E_2)\,\delta(E_2-E_1-\Delta E)$$

$$= \int_{-\infty}^{\infty} \mathcal{W}(E_2-\Delta E, E_2)dE_2$$

$$\simeq \frac{1}{T_{\mathrm{fr}}^2} \int_{-\infty}^{\infty} \exp\left[\frac{\epsilon_2-\Delta E}{T_{\mathrm{fr}}} + \frac{\epsilon_2}{T_{\mathrm{fr}}} - e^{\epsilon_2/T_{\mathrm{fr}}} \right] d\epsilon_2$$

$$= \frac{1}{T_{\mathrm{fr}}} e^{-\Delta E/T_{\mathrm{fr}}} \int_{0}^{\infty} \xi e^{-\xi} d\xi$$

$$= \frac{1}{T_{\mathrm{fr}}} e^{-\Delta E/T_{\mathrm{fr}}}.$$

$$(A18)$$

Of course, this is the result for $\Delta E > 0$, so in general the
probability distribution for the gap reads

$$\mathcal{W}(\Delta E) = \begin{cases} 0 & \text{for } \Delta E < 0, \\ (1/T_{\mathrm{fr}})e^{-\Delta E/T_{\mathrm{fr}}} & \text{for } \Delta E > 0. \end{cases}$$

$$(A19)$$

It might seem surprising at first glance that this probability
decays only as exponential, not the sharper $e^{-\exp \xi}$ function.
Indeed, if we imagine a gap forming as fixing the lowest
energy at the typical place and then looking at the second
lowest level, then the probability of the gap decays in a very
sharp, double-exponential way. However, if we think differ-
ently, that E_2 is fixed in a typical place and then the gap is
determined by where E_1 is then the probability of E_1 going
down is just exponential. Thus, from this argument we gain
the following insight. The probability of the gap is exponen-
tial because the dominant possibility for the gap to be large
has an unusually low E_1 rather than a high E_2. Thus, a large
gap is typically the result of the low ground state.

Averages

Using formula (A10), it is easy to compute the expecta-
tion value of the ground state energy:

$$\langle E_g \rangle = \int_{-\infty}^{\infty} E \mathcal{W}(E)dE$$

$$= E_m + T_{\mathrm{fr}} \int_{-\infty}^{\infty} \xi e^{\xi-\exp \xi} d\xi$$

$$= E_m + \frac{\delta B}{\sqrt{2s}} \underbrace{\int_{0}^{\infty} \ln \eta e^{-\eta} d\eta}_{\Gamma'(1)\approx-0.6},$$

$$(A20)$$

where $\xi = \epsilon/T_{\mathrm{fr}}$ and $\eta = e^{\xi}$. As a fact of mathematical curios-
ity, there appears Γ', the derivative of the Euler Γ
function—a rather infrequent guest in physics calculations.
However exotic mathematically, this term is smaller than the
logarithmic correction term in Eq. (A7) and so it must be
neglected if the first approximation is used for E_m.
We can also compute the average value of the second
lowest energy:

$$\langle E_2 \rangle = \int_{-\infty}^{\infty} E \mathcal{W}_2(E) dE$$

$$= E_m + T_{\text{fr}} \int_{-\infty}^{\infty} \xi e^{2\xi - \exp \xi} d\xi$$

$$= E_m + \frac{\delta B}{\sqrt{2s}} \underbrace{\int_0^{\infty} \eta \ln \eta e^{-\eta} d\eta}_{\Gamma'(2) = 1 + \Gamma'(1) \approx 0.4}. \tag{A21}$$

This is greater than the average ground-state energy (A20) by just the amount $T_{\text{fr}} = \delta B / \sqrt{2s}$.

The latter result is also consistent with the fact that the average gap, according to Eq. (A19) is equal to

$$\langle \Delta E \rangle = T_{\text{fr}} = \delta B / \sqrt{2s}. \tag{A22}$$

[1] S. Bresler and D. Talmud, Dokl. Akad. Nauk SSSR **47**, 326 (1944).
[2] S. Bresler and D. Talmud, Dokl. Akad. Nauk SSSR **47**, 367 (1944).
[3] A. V. Finkelstein and O. B. Ptitsyn, *Protein Physics: A Course of Lectures* (Academic Press, New York, 2002).
[4] J. D. Bryngelson and P. G. Wolynes, Proc. Natl. Acad. Sci. U.S.A. **84**, 7524 (1987).
[5] E. I. Shakhnovich and A. M. Gutin, Biophys. Chem. **34**, 187 (1989).
[6] B. Derrida, Phys. Rev. Lett. **45**, 79 (1980).
[7] C. D. Sfatos and E. I. Shakhnovich, Phys. Rep. **288**, 77 (1997).
[8] M. Oliveberg and E. I. Shakhnovich, Curr. Opin. Struct. Biol. **16**, 68 (2006).
[9] P. L. Geissler, E. I. Shakhnovich, and A. Y. Grosberg, Phys. Rev. E **70**, 021802 (2004).
[10] A. R. Khokhlov and P. G. Khalatur, Phys. Rev. Lett. **82**, 3456 (1999).
[11] A. R. Khokhlov and P. G. Khalatur, Curr. Opin. Solid State Mater. Sci. **8**, 3 (2004).
[12] V. S. Pande, A. Y. Grosberg, and T. Tanaka, Biophys. J. **73**, 3192 (1997).
[13] H. Li, R. Helling, C. Tang, and N. Wingreen, Science **273**, 666 (1996).
[14] C. D. Sfatos, A. M. Gutin, and E. I. Shakhnovich, Phys. Rev. E **51**, 4727 (1995).
[15] K. A. Dill and D. Stigter, Adv. Protein Chem. **46**, 59 (1995).
[16] T. Garel, L. Leibler, and H. Orland, J. Phys. II **4**, 2139 (1994).
[17] S. S. Plotkin and J. N. Onuchic, Q. Rev. Biophys. **35**, 111 (2002).
[18] L. Reagan and W. F. Degrado, Science **241**, 976 (1989).
[19] A. Amatori, G. Tiana, J. F-Borg, A. Trovato, and R. A. Broglia, J. Chem. Phys. **123**, 054904 (2005).
[20] A. Amatori, J. Fenkinghoff-Borg, G. Tiana, and R. A. Broglia, Phys. Rev. E **73**, 061905 (2006).
[21] V. S. Pande, A. Y. Grosberg, and T. Tanaka, Rev. Mod. Phys. **72**, 259 (2000).
[22] J. Bouchaud and M. Mezard, J. Phys. A **30**, 7997 (1997).
[23] K. A. Dill *et al.*, Protein Sci. **4**, 561 (1995).
[24] P. M. Chaikin and T. C. Lubensky, *Principles of Condensed Matter Physics* (Cambridge University Press, Cambridge, England, 1995).
[25] P. Wong and D. Frishman, PLOS Comput. Biol. **2**, 392 (2006).